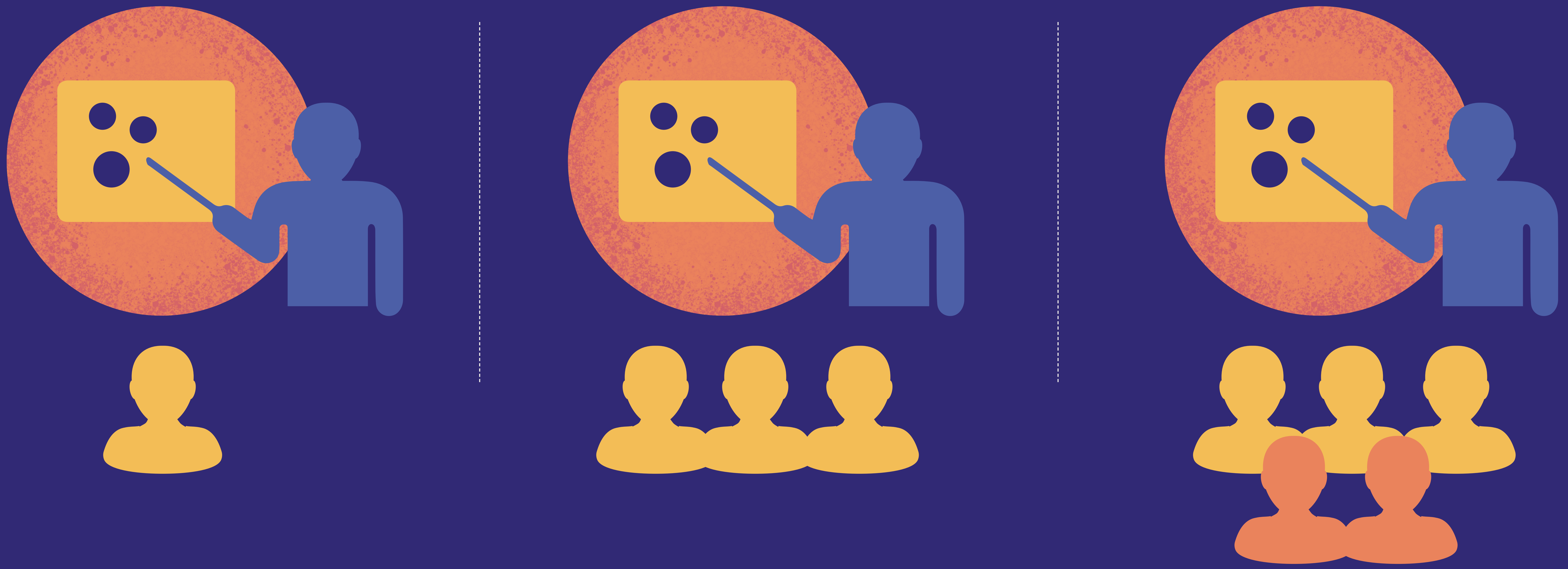
Equity and Ethics in Data Products

Step by step processes to avoid racism, sexism, homophobia and more in data and analysis.

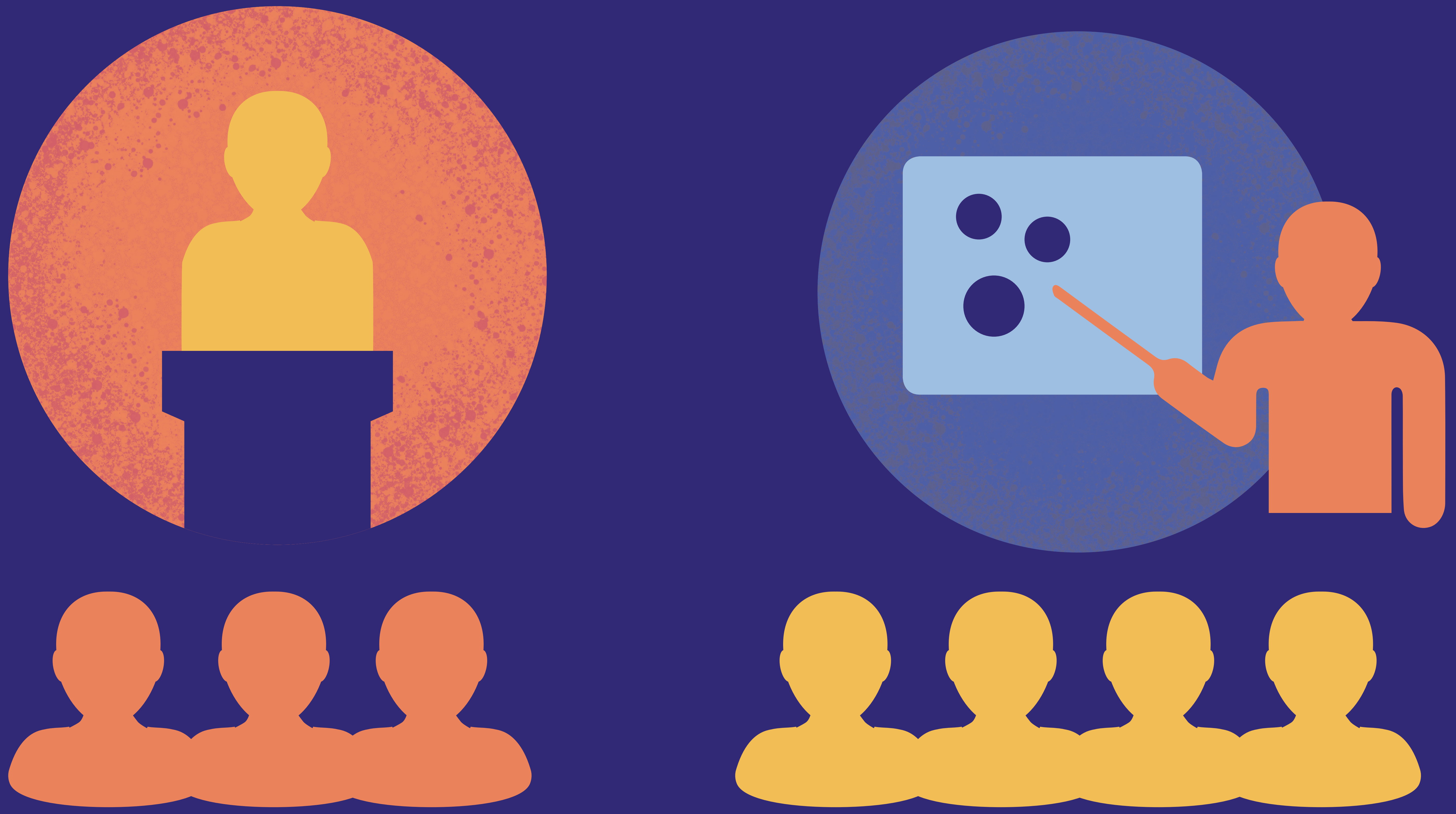
What is the average size classroom?

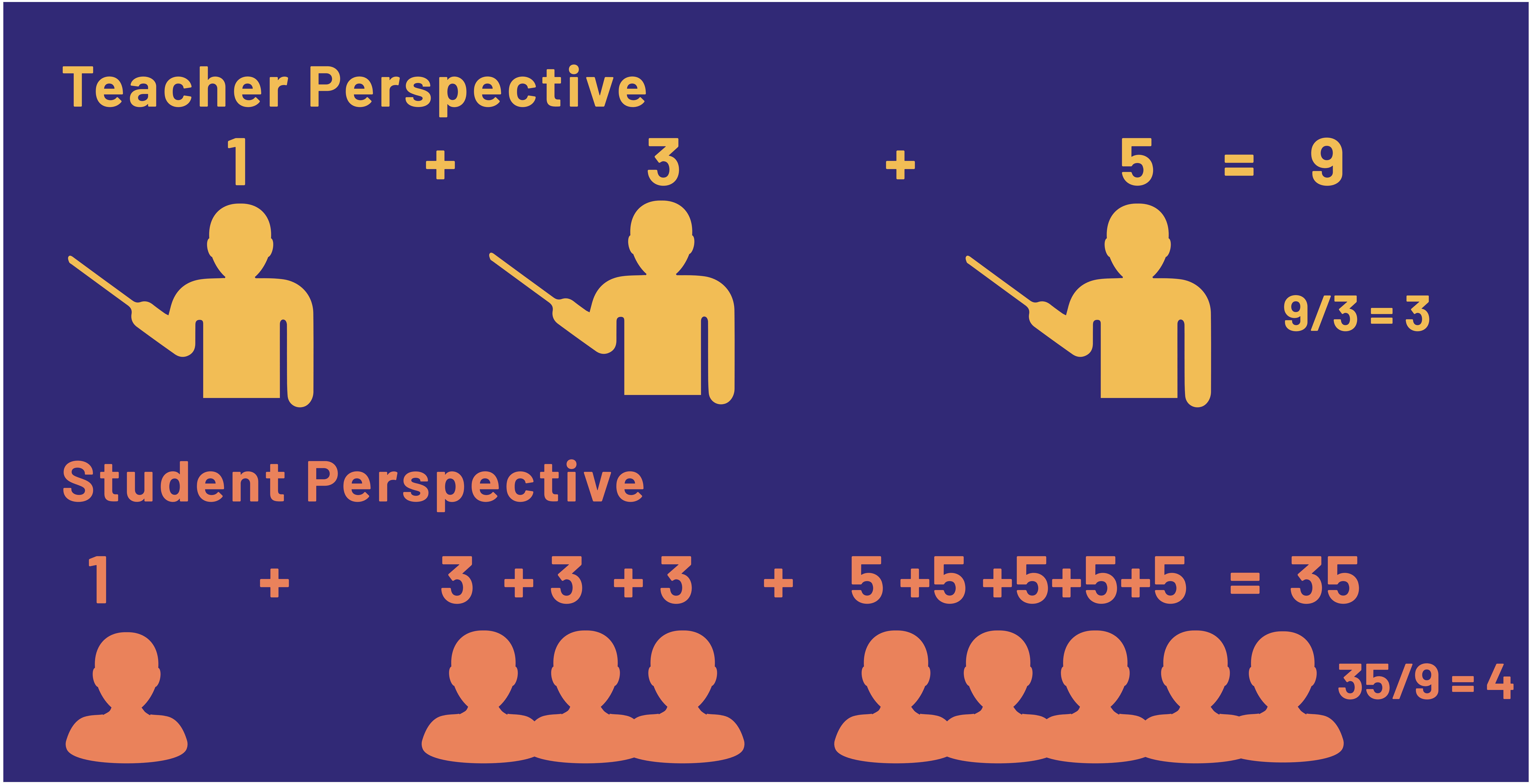


The average classroom is 3 students per class.

The average classroom is 4 students per class.

Both are correct.





Teacher Perspective

1 + 3 + 5 = 9

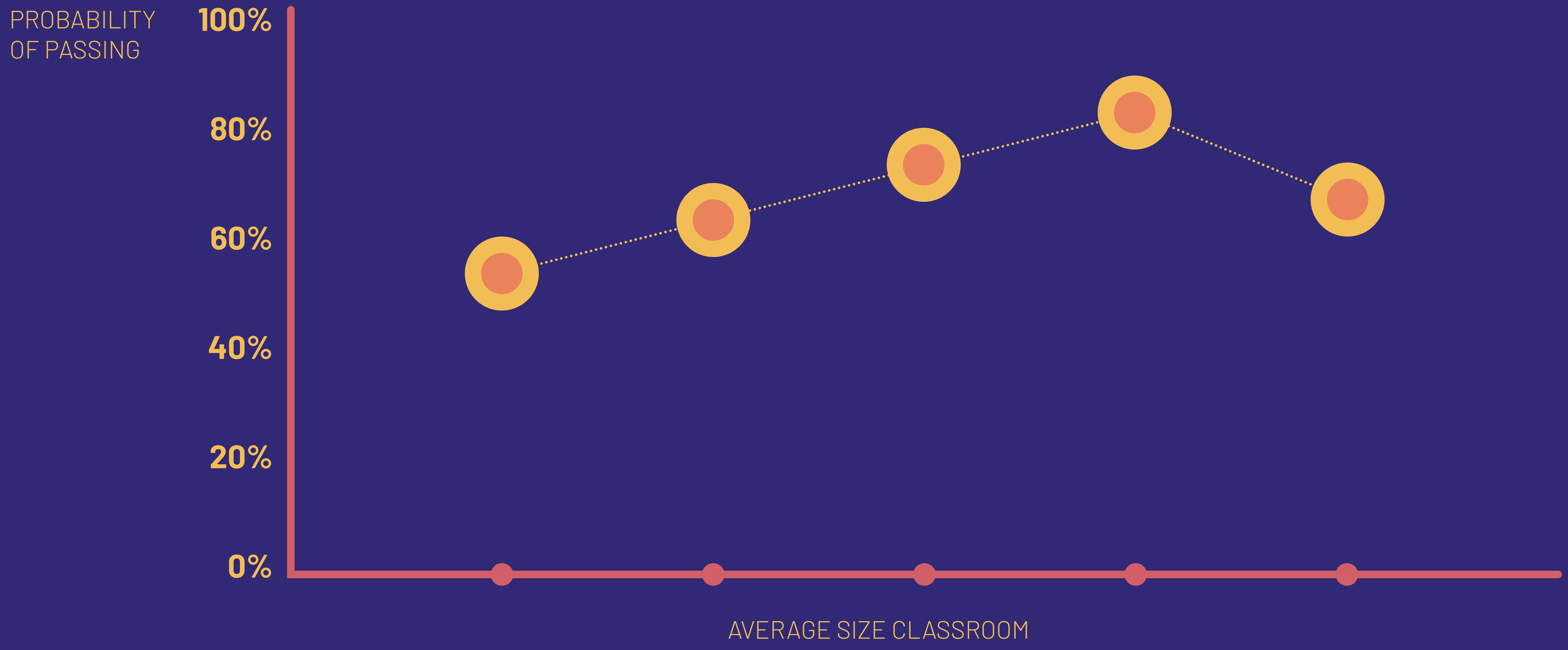
9/3 = 3

Student Perspective

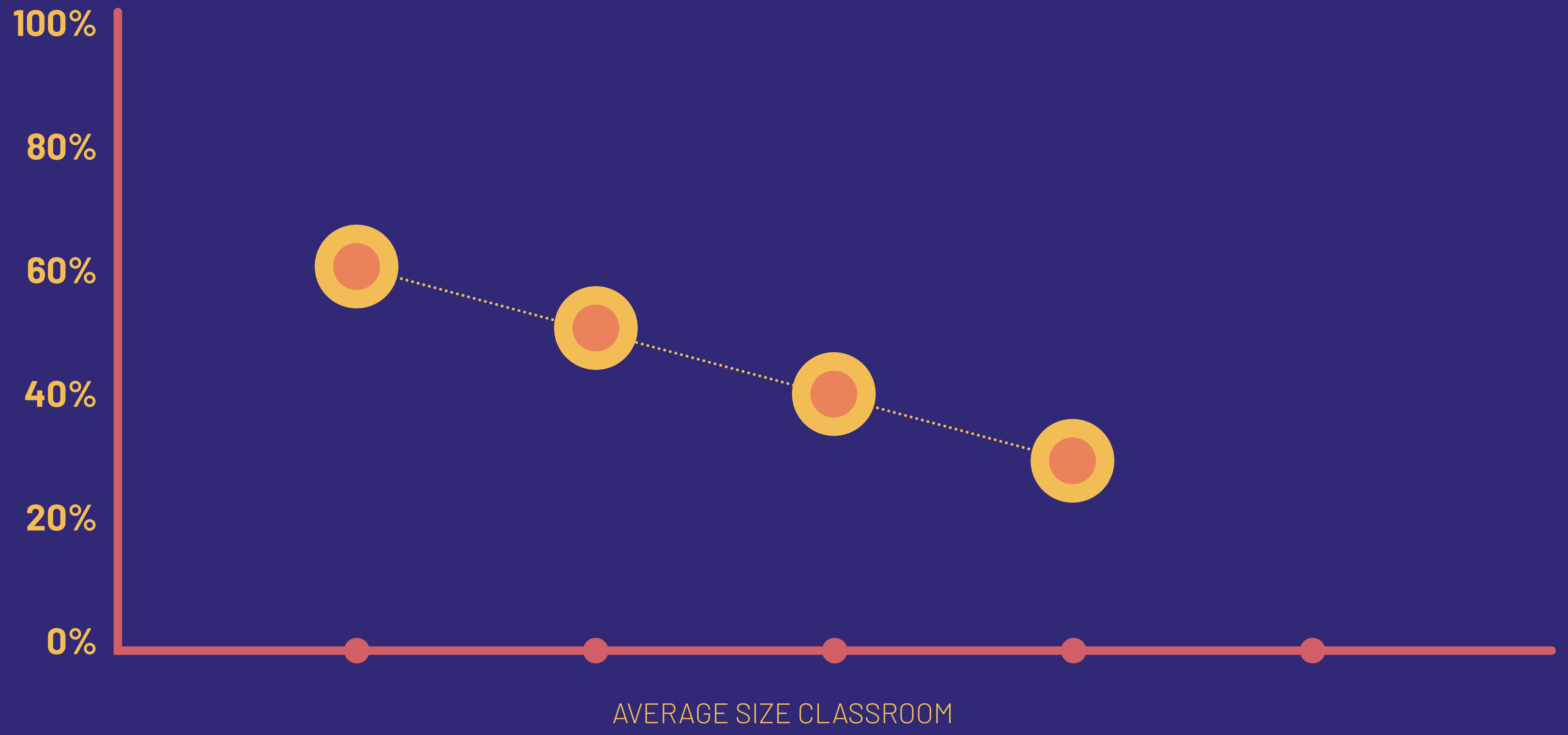
1 + 3 + 3 + 3 + 5 +5 +5+5+5 = 35

35/9 = 4

Here is the relationship between class size and academic performance from the student's POV.



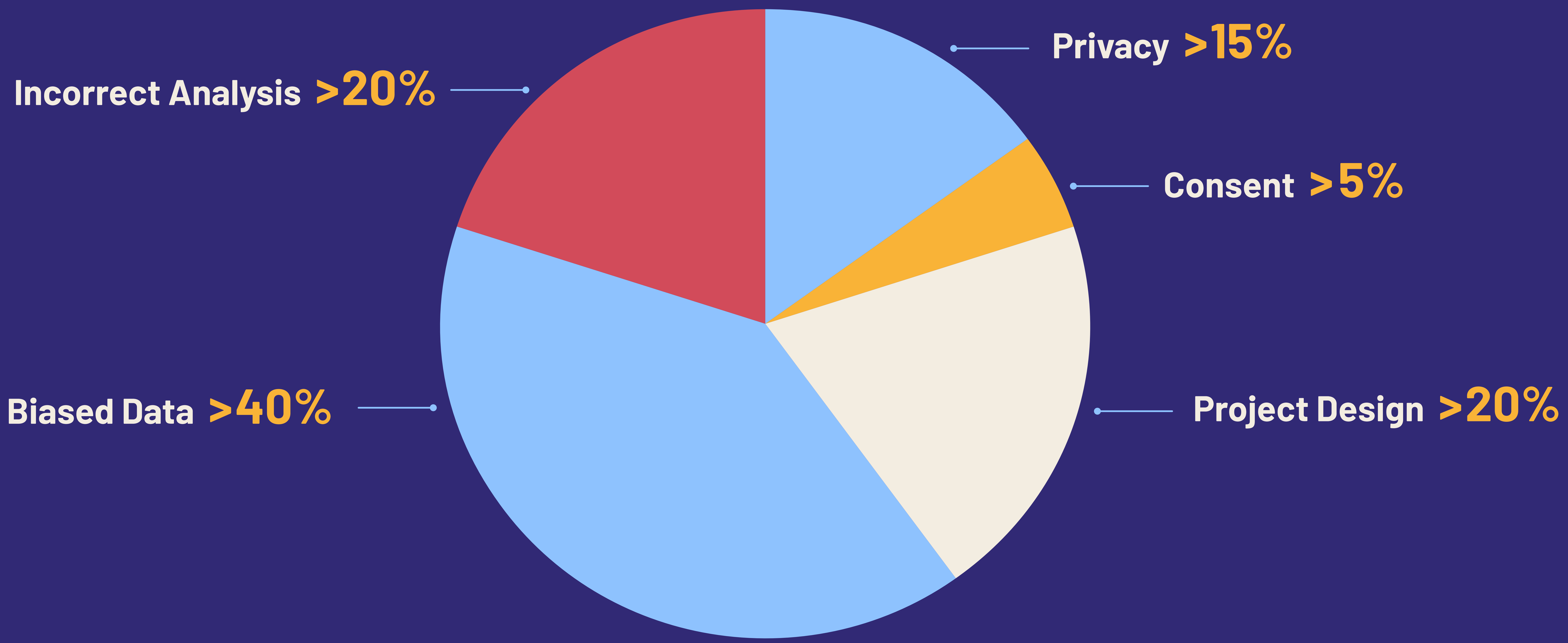
Here is the relationship between class size and academic performance from the teacher’s POV.



WE ALL COUNT

project for equity in data science

Data Ethics = **Privacy?**



Organizations tend to take a legalistic approach to data ethics - what are the laws, privacy, disclosure requirements, etc.

Consumers and citizens are taking a much broader and less legalistic attitude towards data ethics.

Sources of bias can be identified in each step of the data life cycle.

Funding

Motivation

Project Design

Data Collection & Sourcing

Analysis

Interpretation

Comunication & Distribution

We All Count Tools

We All Count believes that the world is a little too full of people pointing out problems without offering solutions. WAC is committed to providing practical resources to help anyone who wants to make their data science more equitable.

Funding

70% GREATER

An analysis of 30 years of educational research by scholars at Johns Hopkins University found that when a maker of an educational intervention conducted its own research or paid someone to do the research, the results commonly showed greater benefits for students than when the research was independent. **On average, the developer research showed benefits — usually improvements in test scores — that were 70 percent greater than what independent studies found.**

Motivation

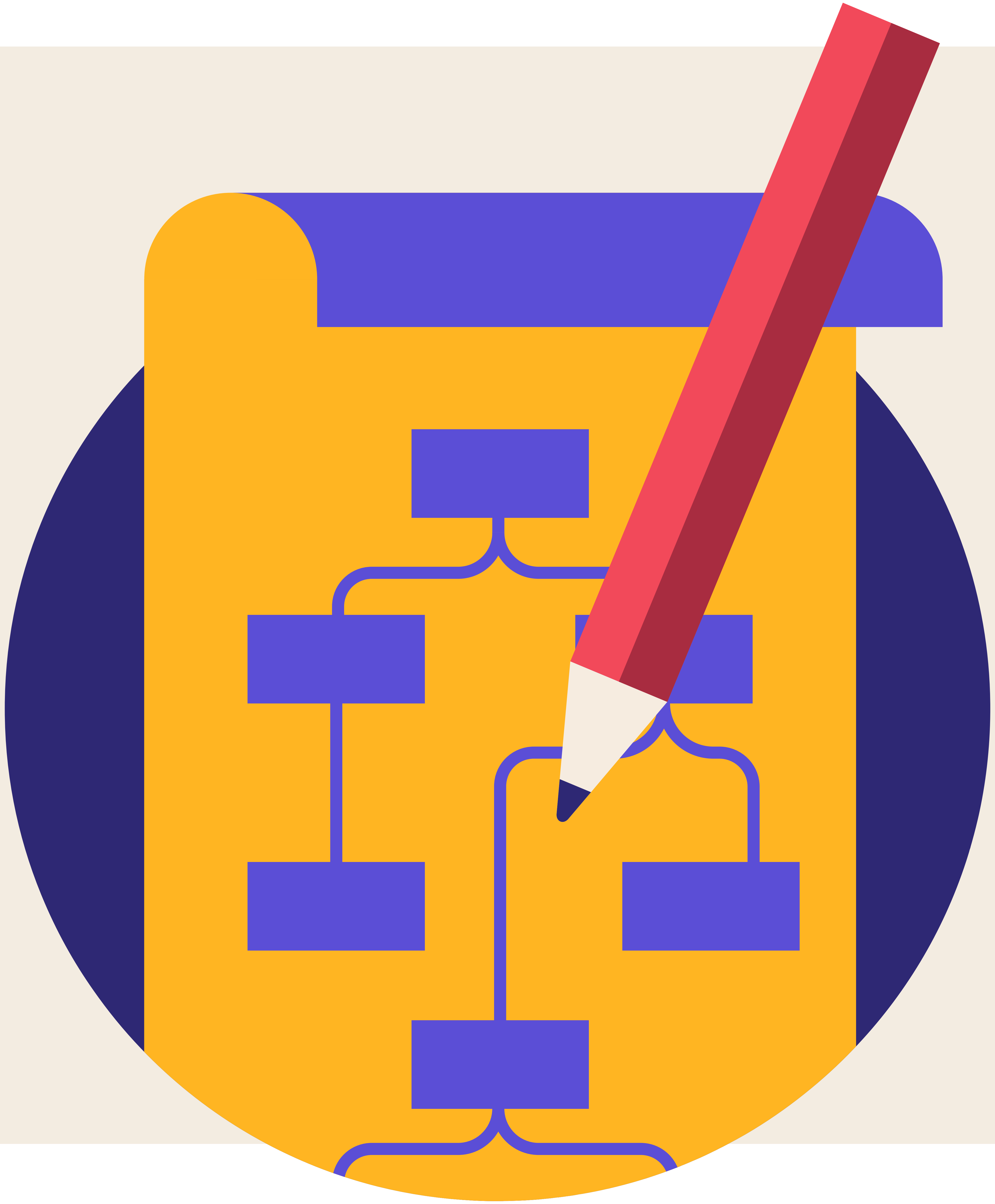
NYC & Rats

Why is this data project being done?

Tension between explicit purpose (understand if this works) and implicit purpose (have a good report for the annual general meeting).

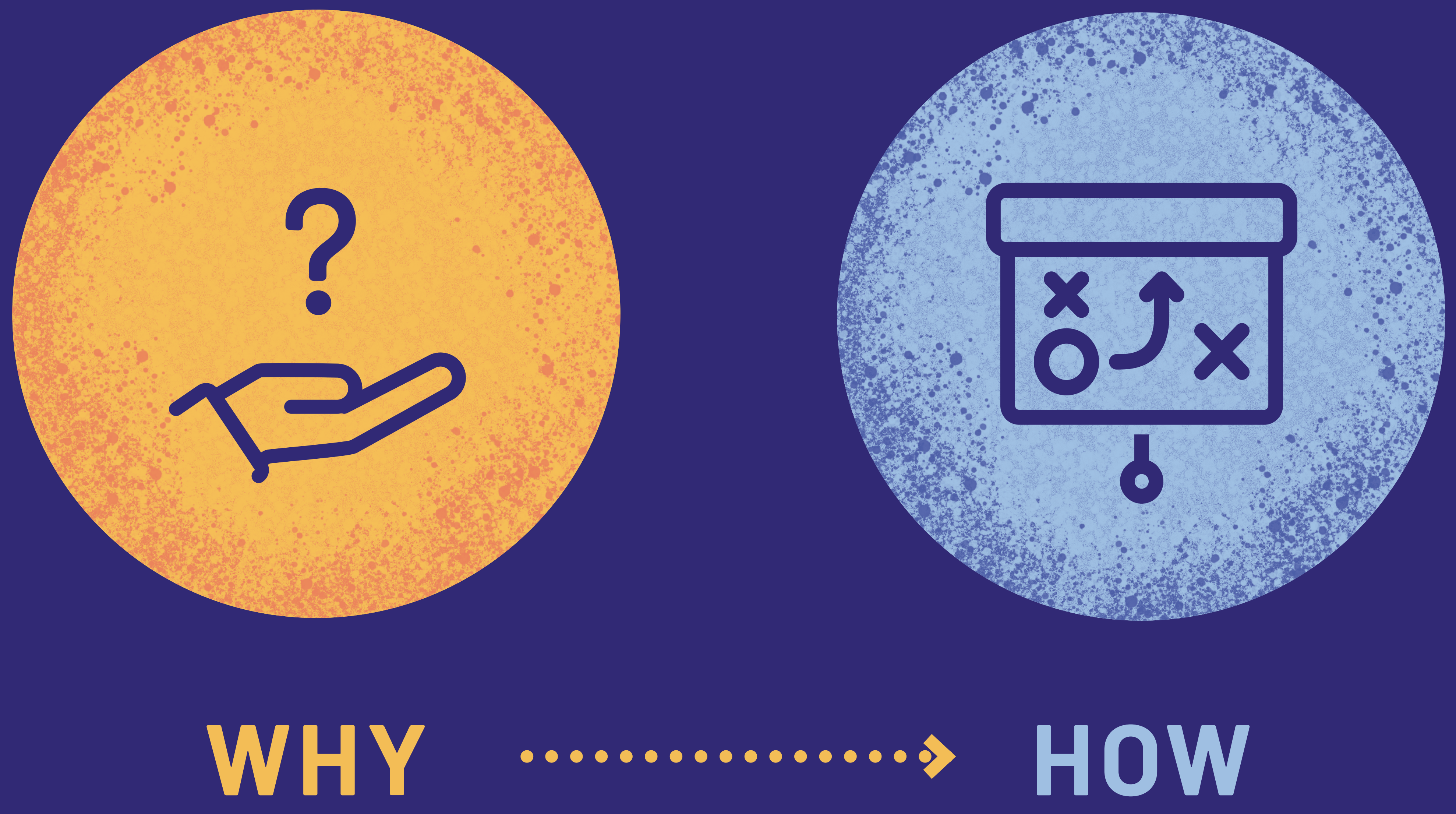
Motivation Statement

Project Design



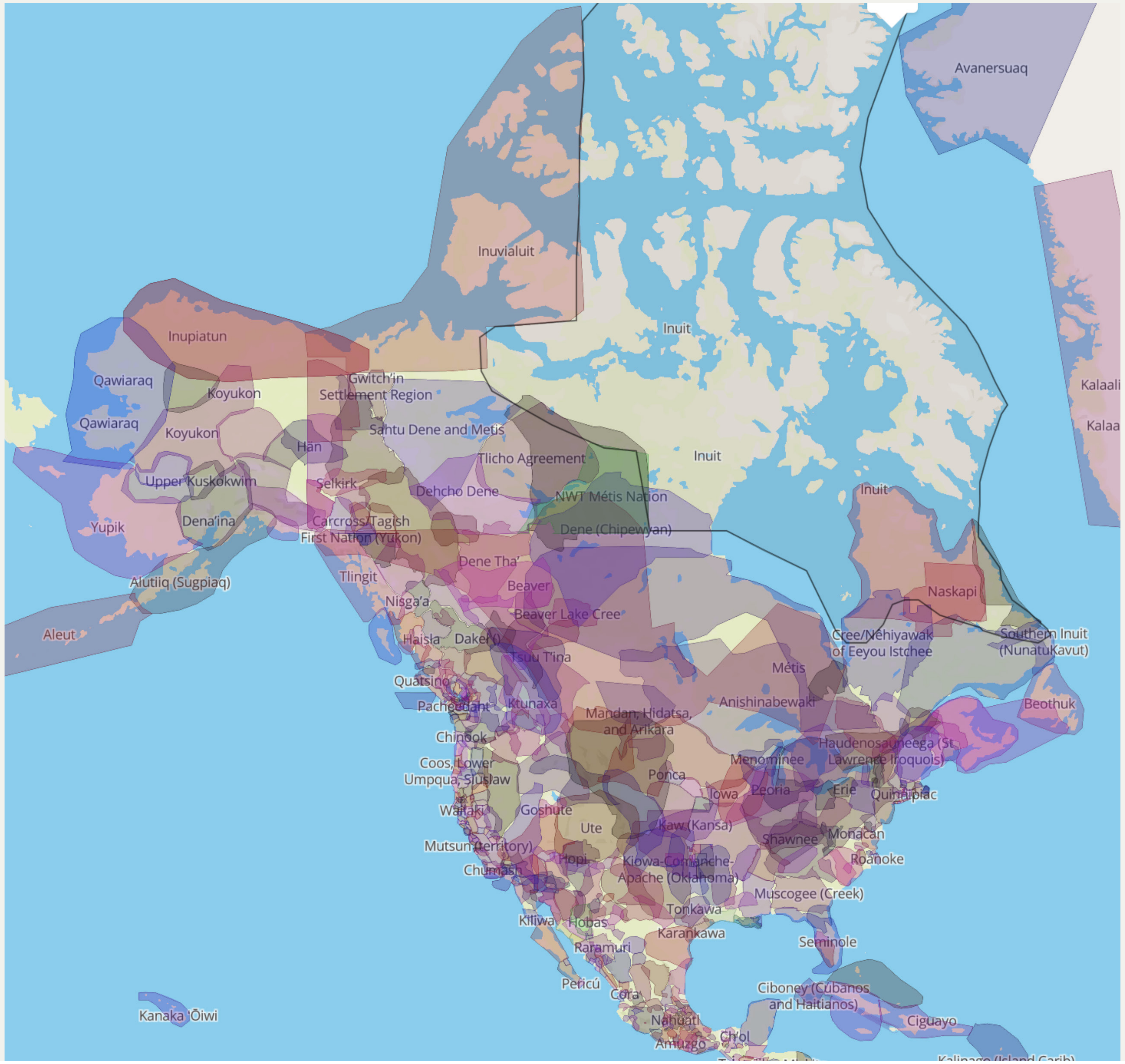
Project Design is the phase where the WHY becomes the HOW

Critical step in data equity

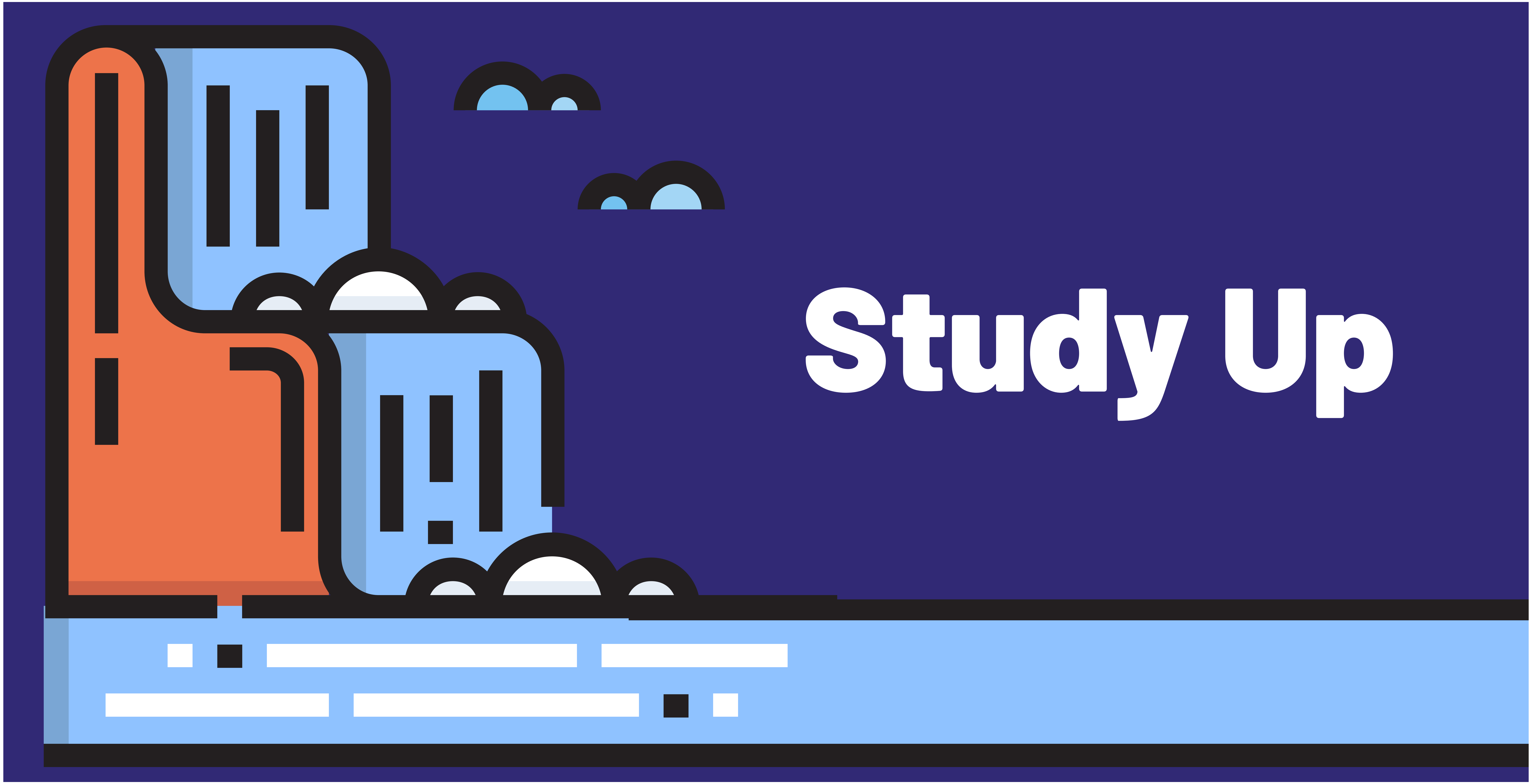


WHY → HOW

Sample design based on definitions - whose definitions?



RCTs:
The Gold Standard … of *What?*

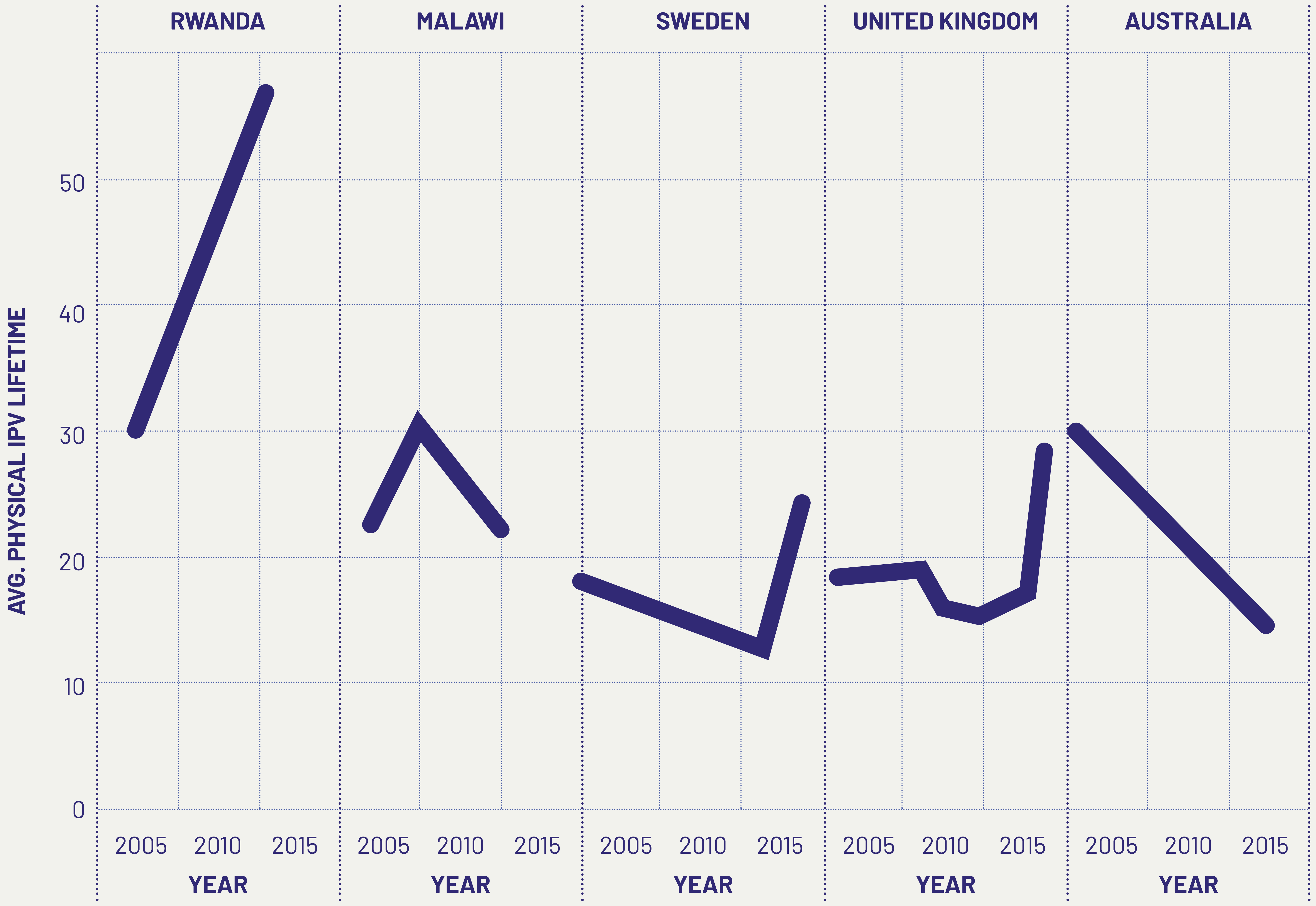


Study Up

Data Collection & Sourcing

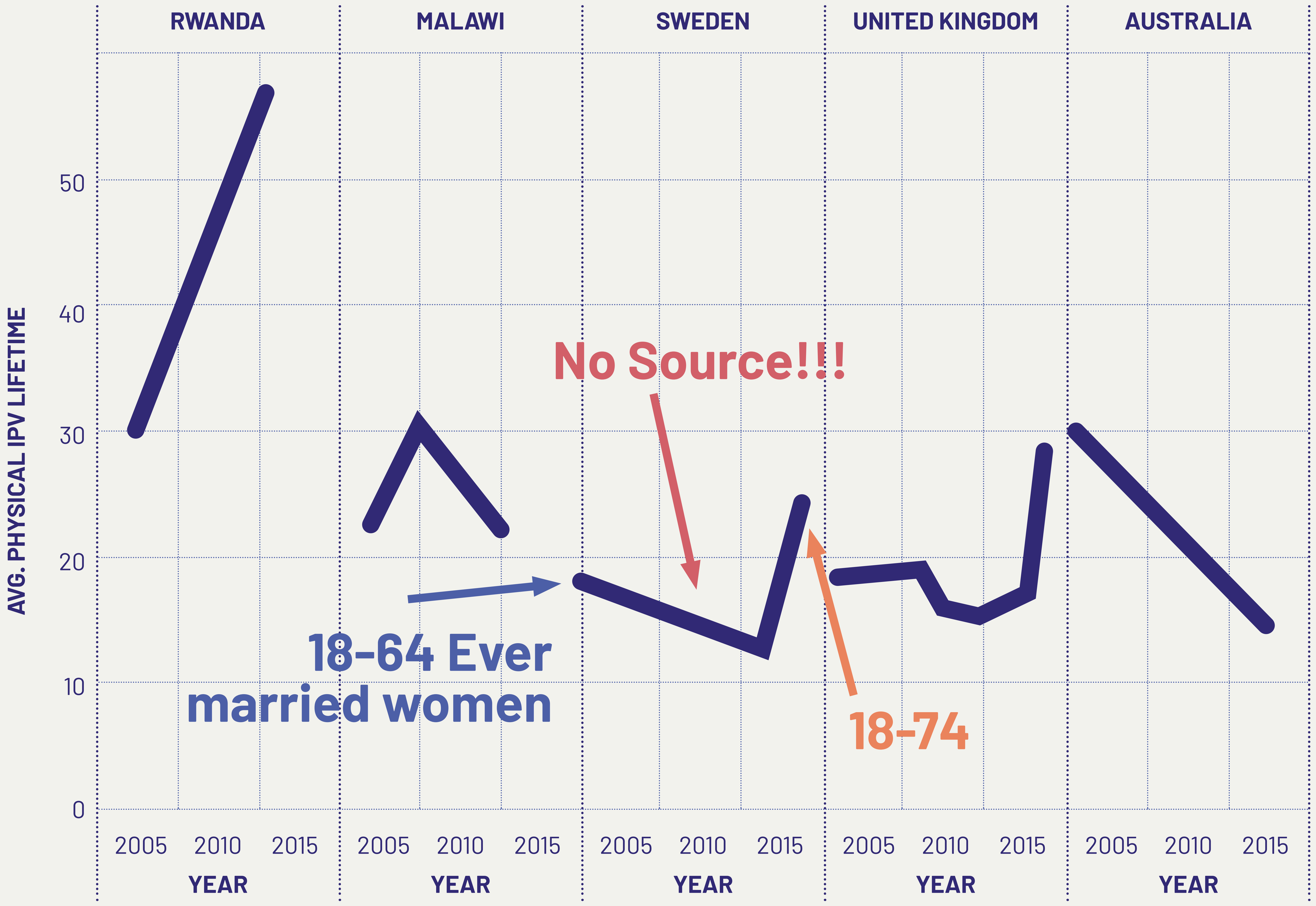
Looking at trends in violence interpersonal violence within countries over time

A stunning amount of change in a short amount of time



Creating a data biography for each data point is time consuming in reverse.

Needs to be included with each data product



Data Biographies at the bare minimum must accompany each dataset you are using:

When

What

Who

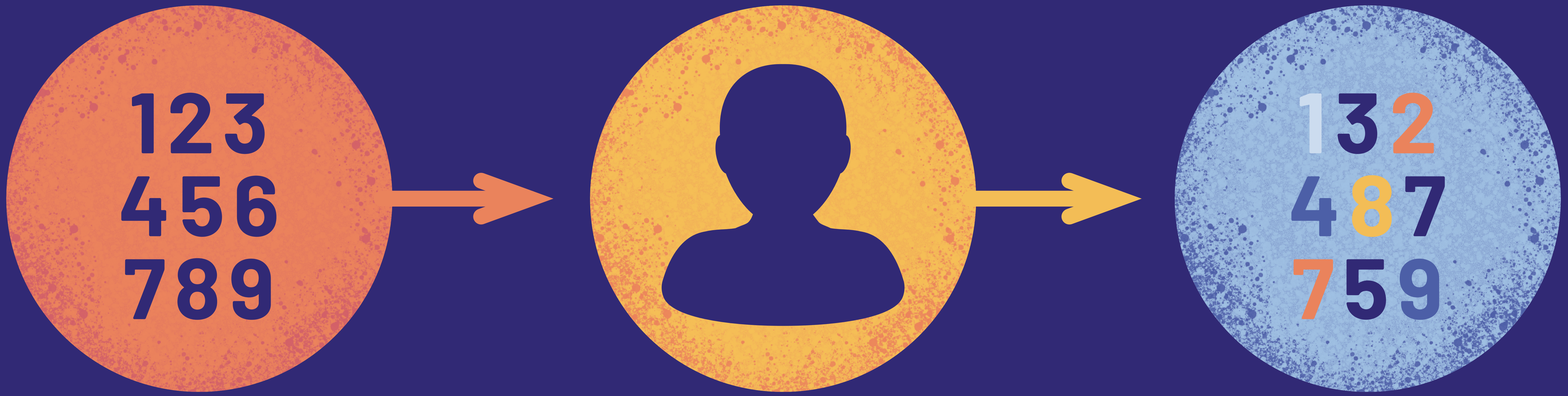
Why

How

Where

Who collects the data matters.

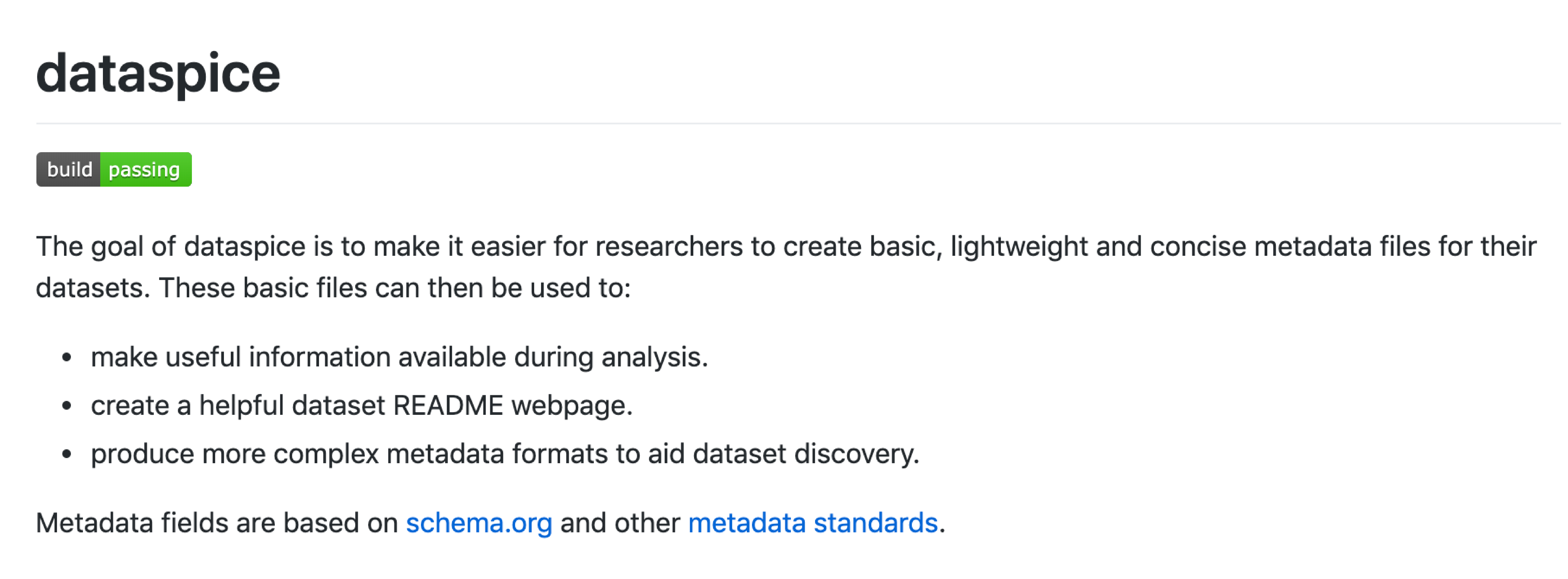
Data collected by higher status enumerators results in different results.



Data Sheets for Data Sets from collab with Google, Georgia Tech, Cornell, Microsoft, Univeristy of Maryland, AI Now

Dataspice from R OpenSci

Data Statements for NLP Emily M. Bender and Batya Friedman



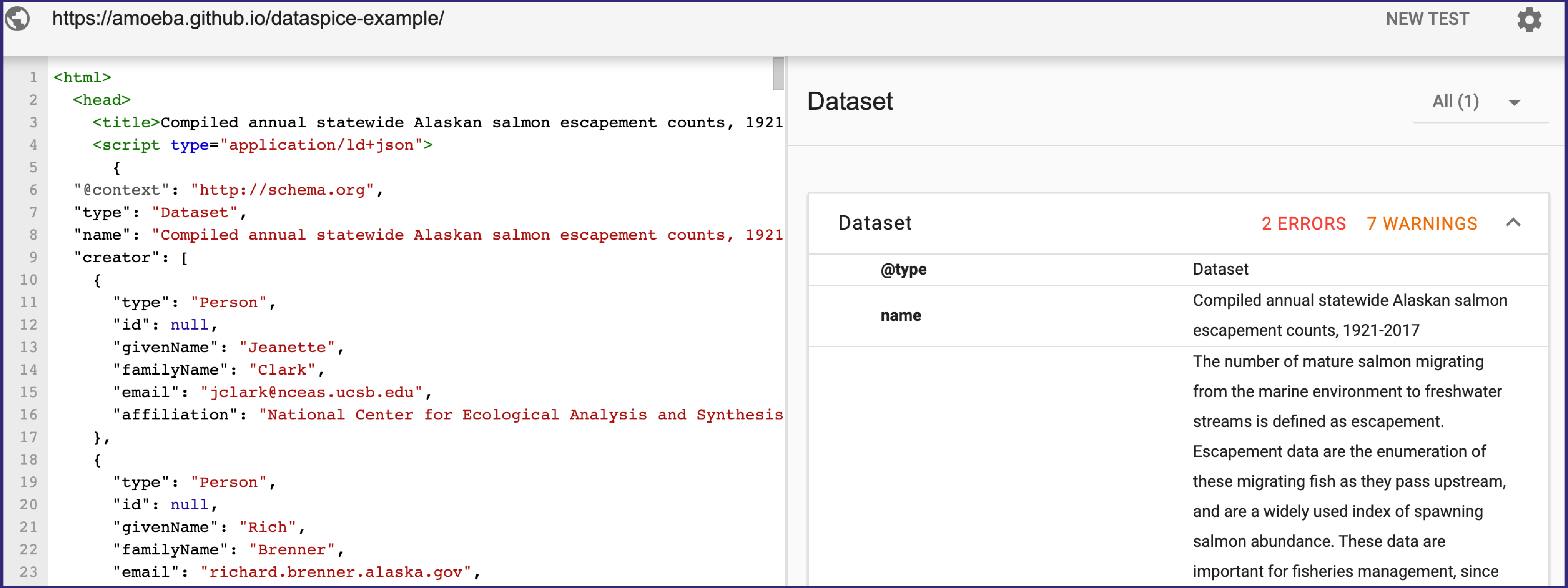
dataspice

build passing

The goal of dataspice is to make it easier for researchers to create basic, lightweight and concise metadata files for their datasets. These basic files can then be used to:

- make useful information available during analysis.
- create a helpful dataset README webpage.
- produce more complex metadata formats to aid dataset discovery.

Metadata fields are based on schema.org and other metadata standards.



https://amoeba.github.io/dataspice-example/

```
<html>
  <head>
    <title>Compiled annual statewide Alaskan salmon escapement counts, 1921
    <script type="application/ld+json">
      {
  "@context": "http://schema.org",
  "type": "Dataset",
  "name": "Compiled annual statewide Alaskan salmon escapement counts, 1921
  "creator": [
    {
      "type": "Person",
      "id": null,
      "givenName": "Jeanette",
      "familyName": "Clark",
      "email": "jclark@nceas.ucsb.edu",
      "affiliation": "National Center for Ecological Analysis and Synthesis
    },
    {
      "type": "Person",
      "id": null,
      "givenName": "Rich",
      "familyName": "Brenner",
      "email": "richard.brenner.alaska.gov",
```

Dataset

Dataset 2 ERRORS 7 WARNINGS

@type	Dataset
name	Compiled annual statewide Alaskan salmon escapement counts, 1921-2017
	The number of mature salmon migrating from the marine environment to freshwater streams is defined as escapement. Escapement data are the enumeration of these migrating fish as they pass upstream, and are a widely used index of spawning salmon abundance. These data are important for fisheries management, since

Data Statements for NLP

Emily M. Bender and Batya Friedman

A. Curation rationale
B. Language variety
C. Speaker demographic
D. Annotator demographic
E. Speech situation
F. Text characteristics
G. Recording quality
H. Other
I. Provenance appendix

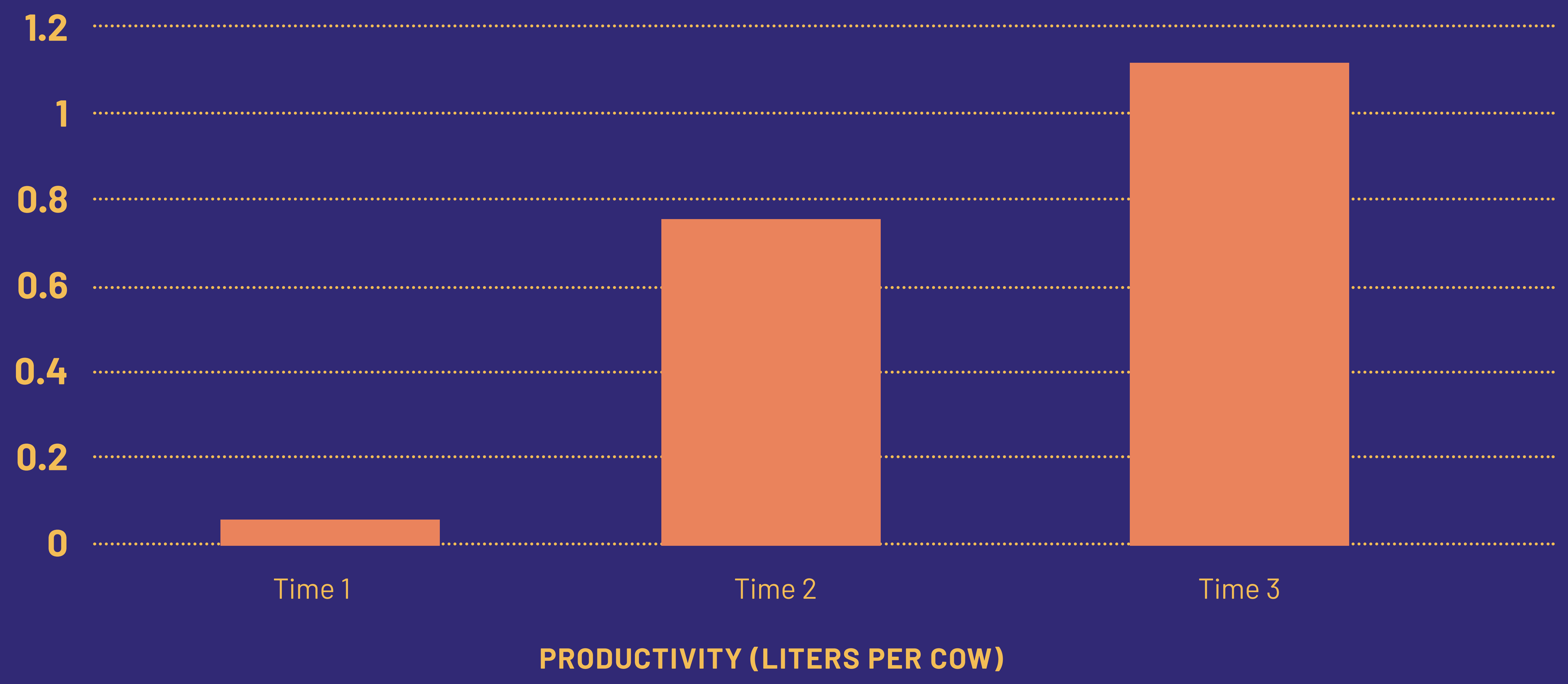
Analysis

Methods Matter

A LOT.

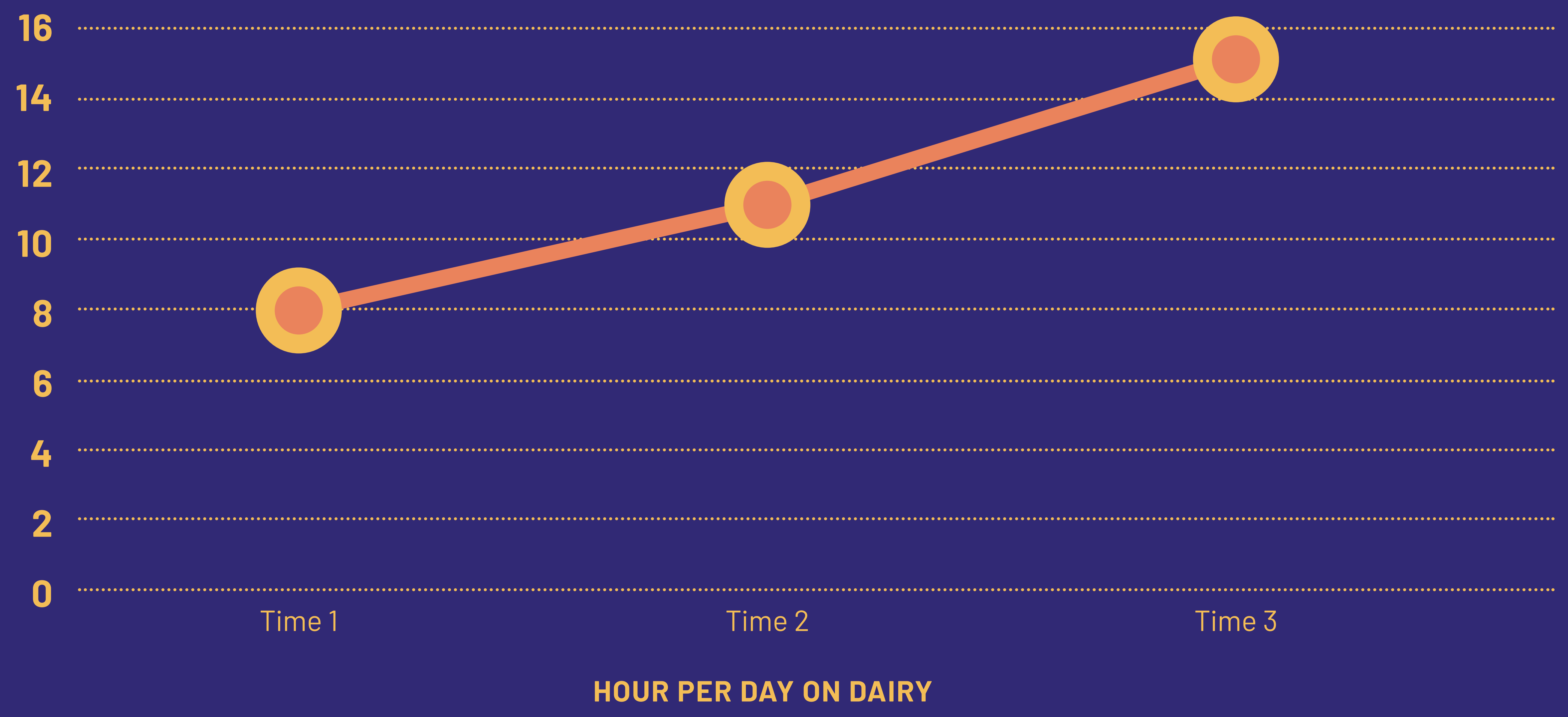
Your world view determines how you measure success.

Productivity increases over time



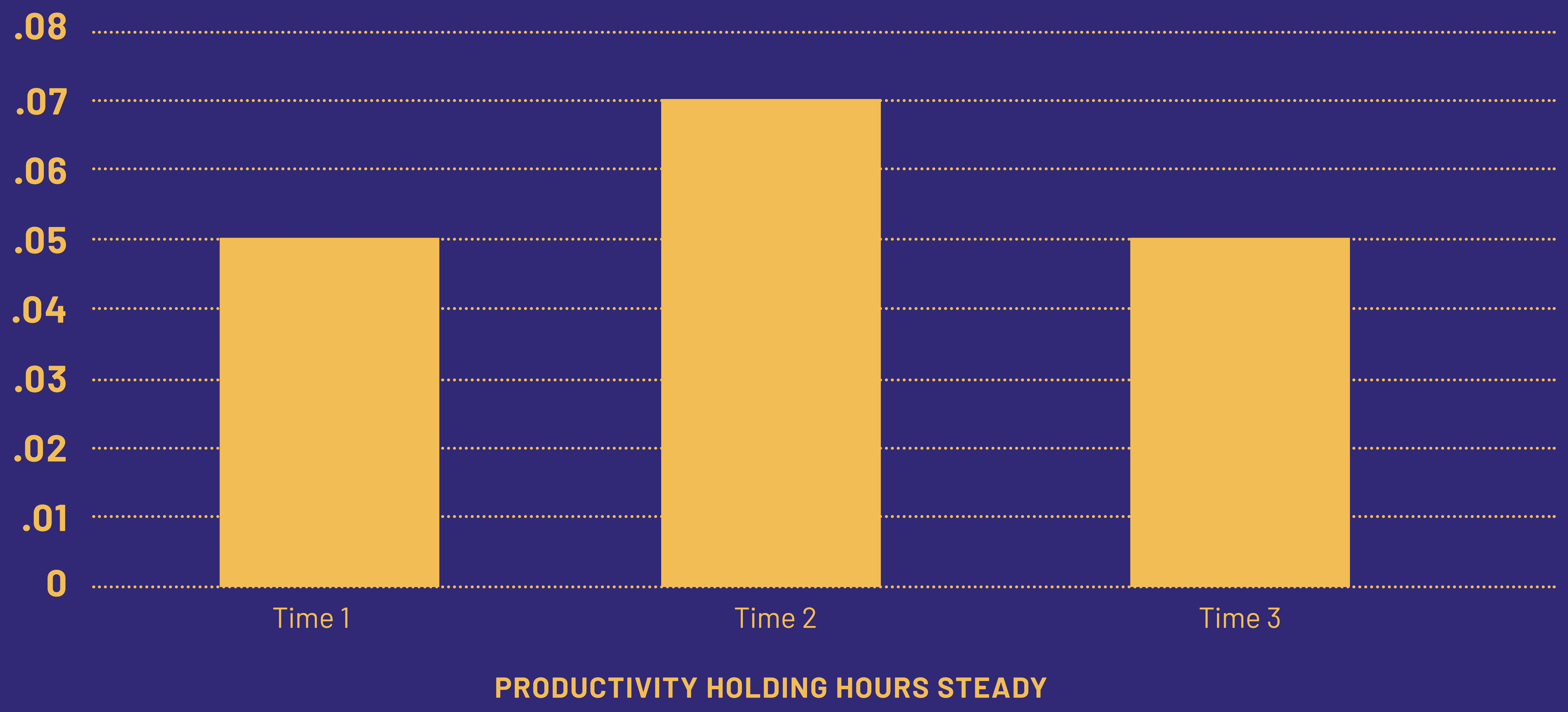
Your world view determines how you measure success.

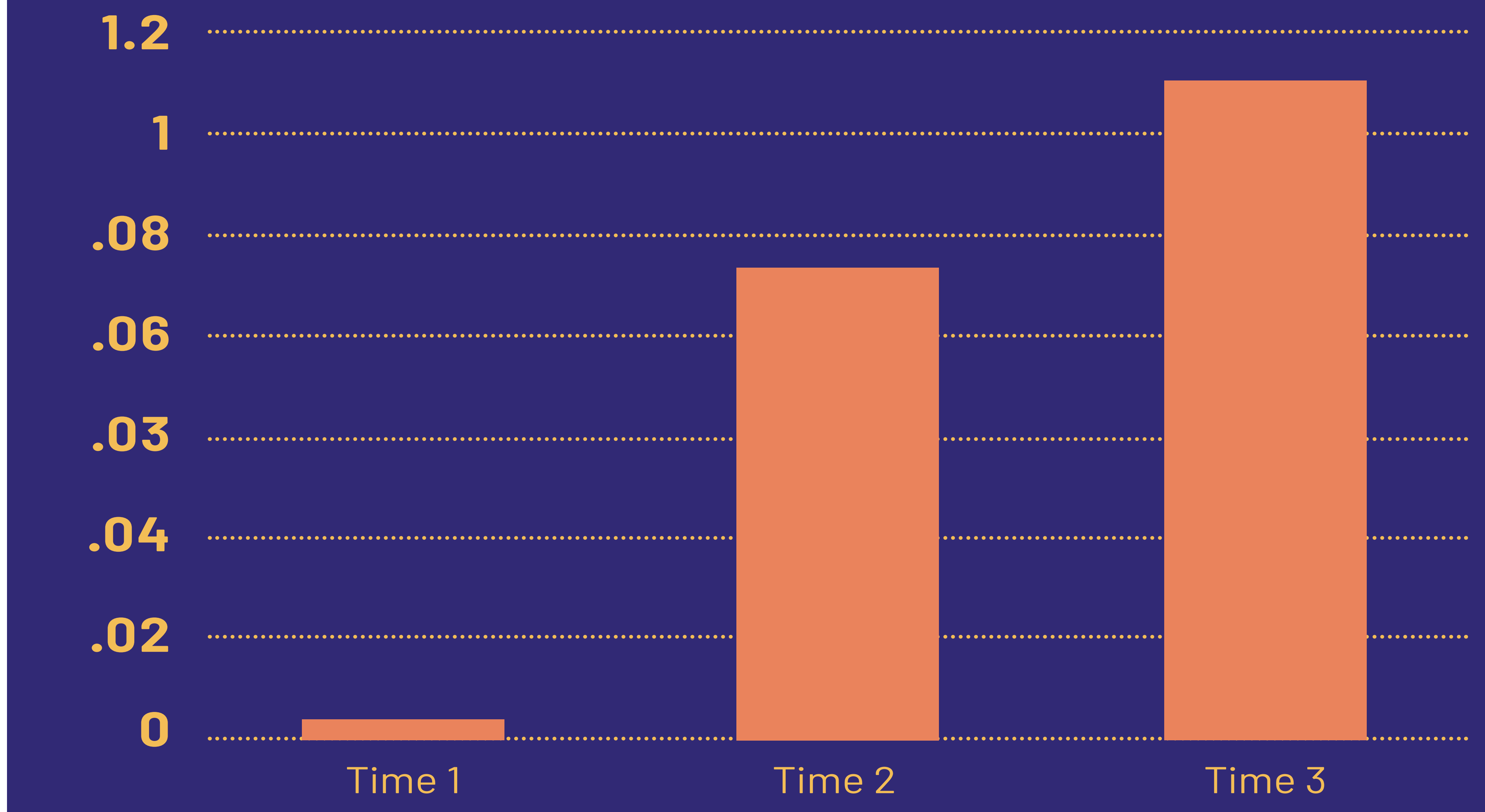
Hours of farm work increases over time



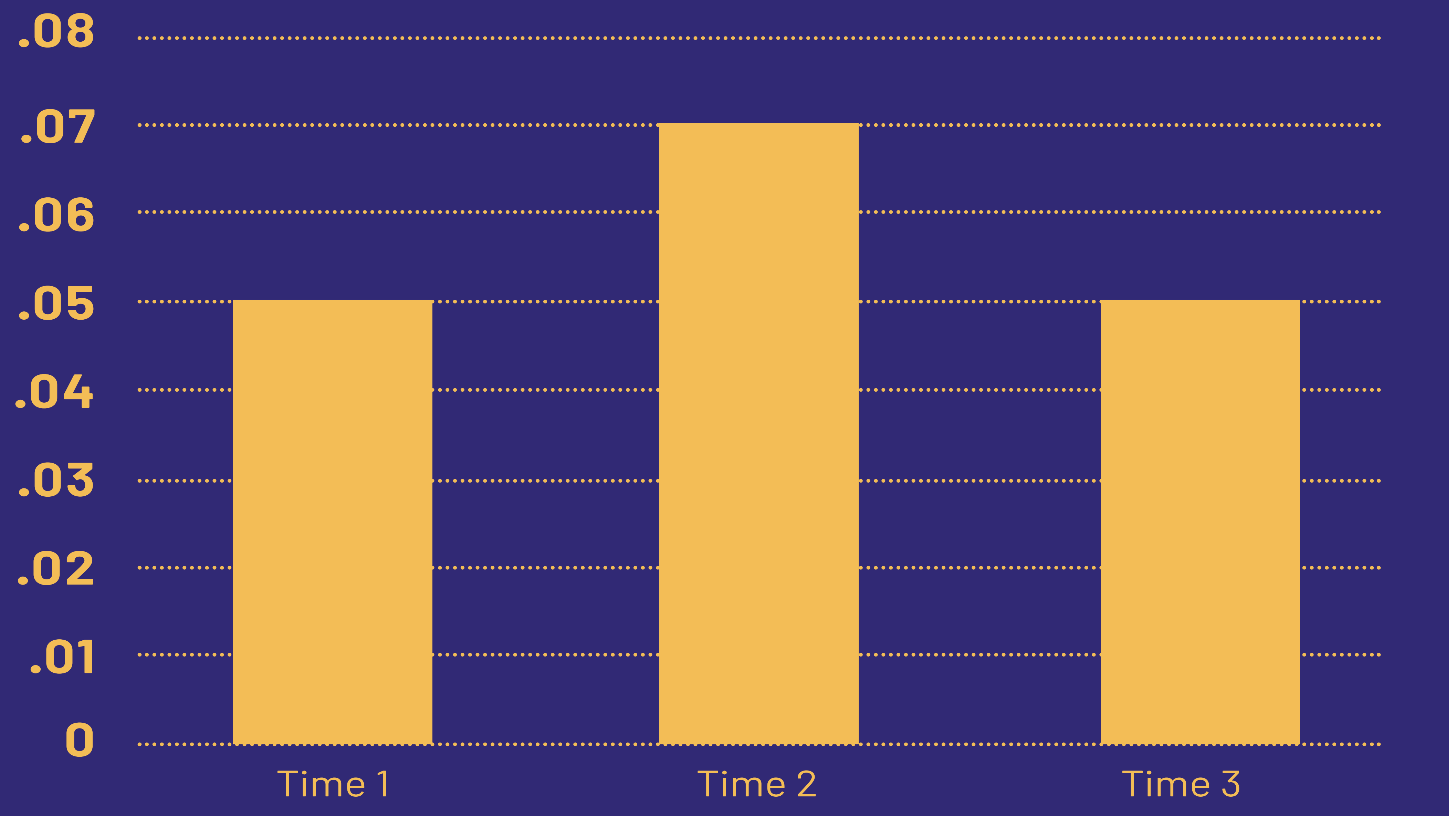
Your world view determines how you measure success.

Productivity controlling for increase in work time





PRODUCTIVITY HOLDING HOURS STEADY



PRODUCTIVITY HOLDING HOURS STEADY

What statistical method you use is based on your world view.

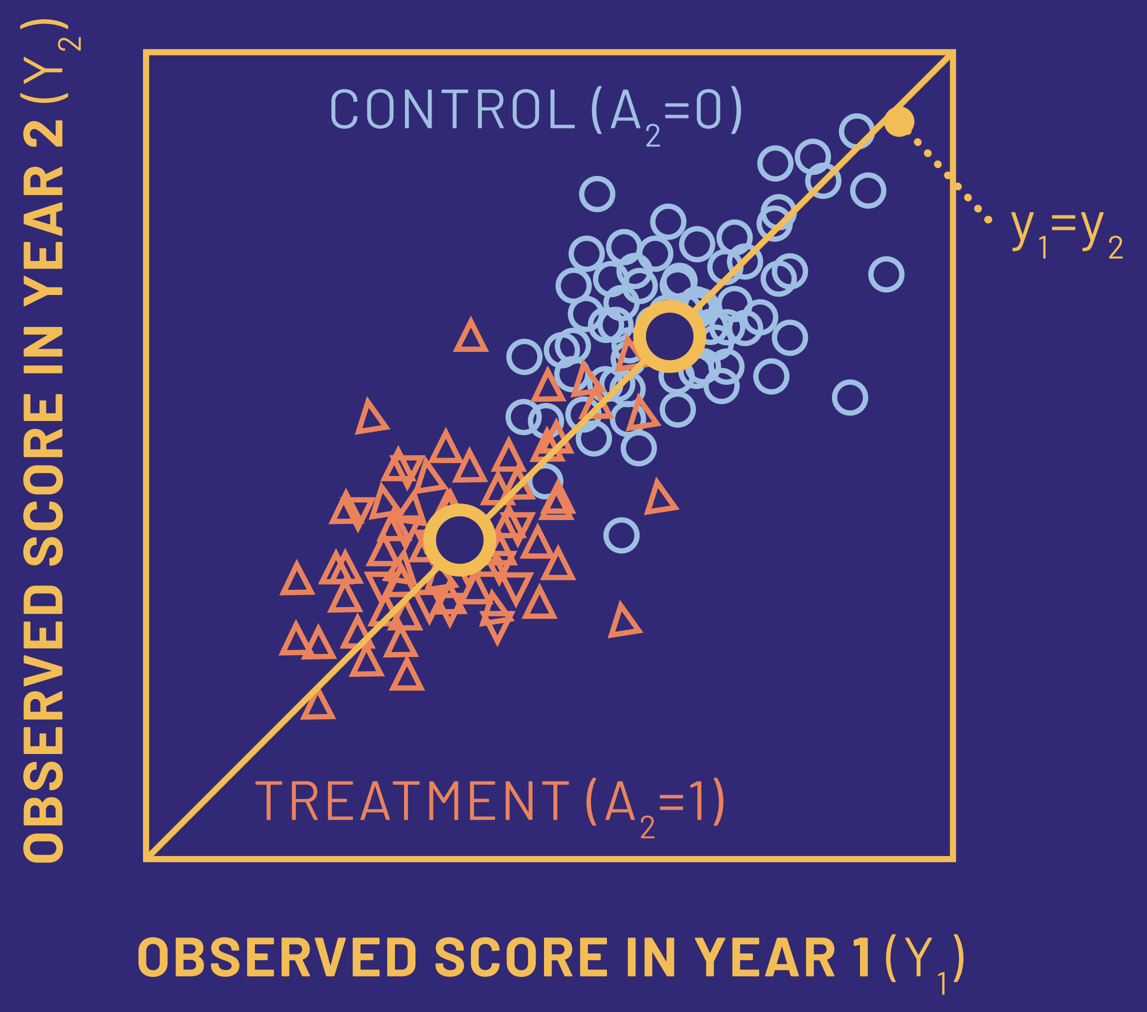
- Special education intervention to help vulnerable kids read better.
- Tested at beginning of year and end of year.
- Assign the treatment to kids who are more likely to need help - since funds are limited.

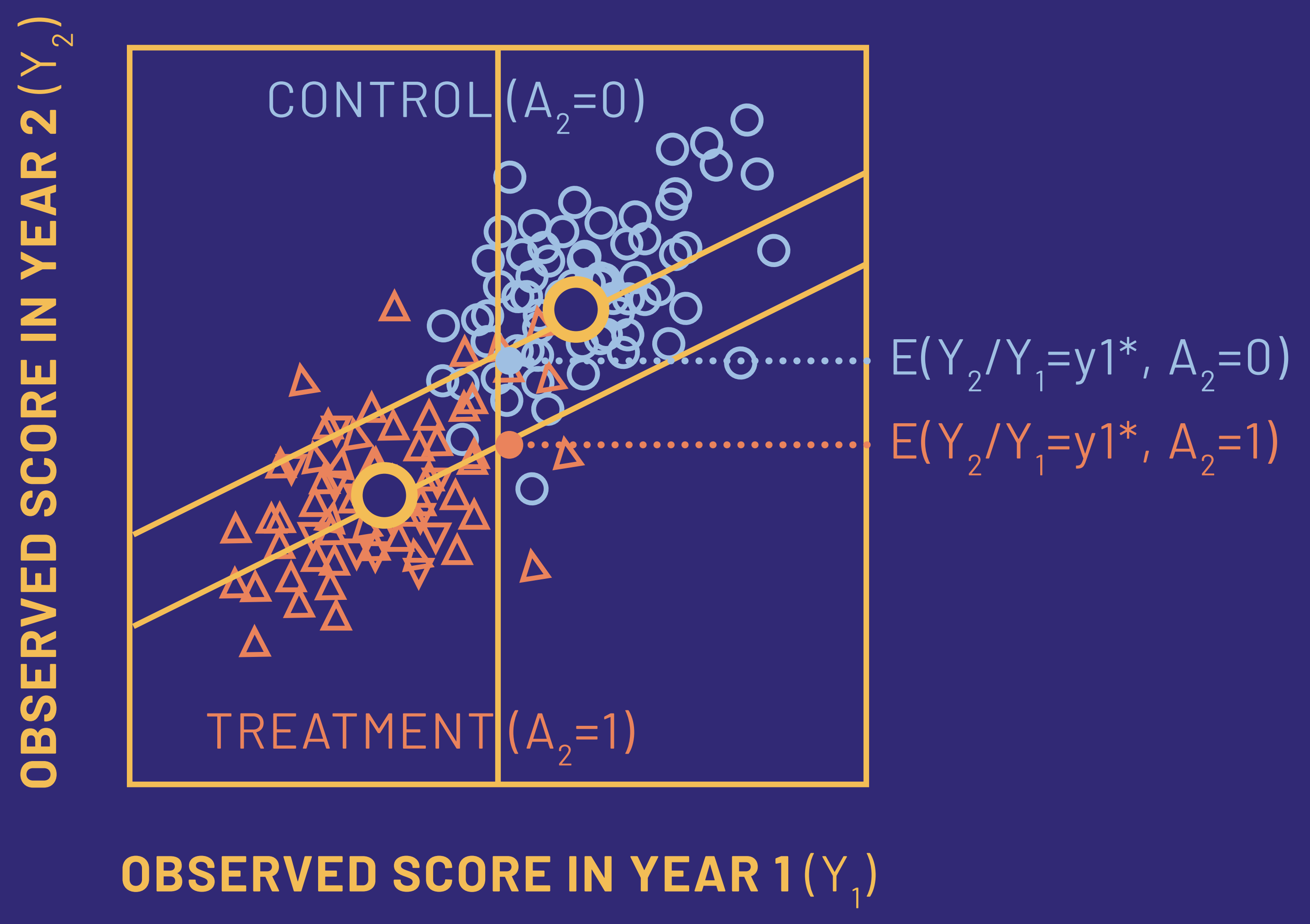
Two analyses done.

One difference in differences, one regression with baseline as covariate.

Analysis #1 concludes that the intervention has no average effect on student reading performance.

Analysis #2 concludes that the intervention has a large negative effect on student reading performance.

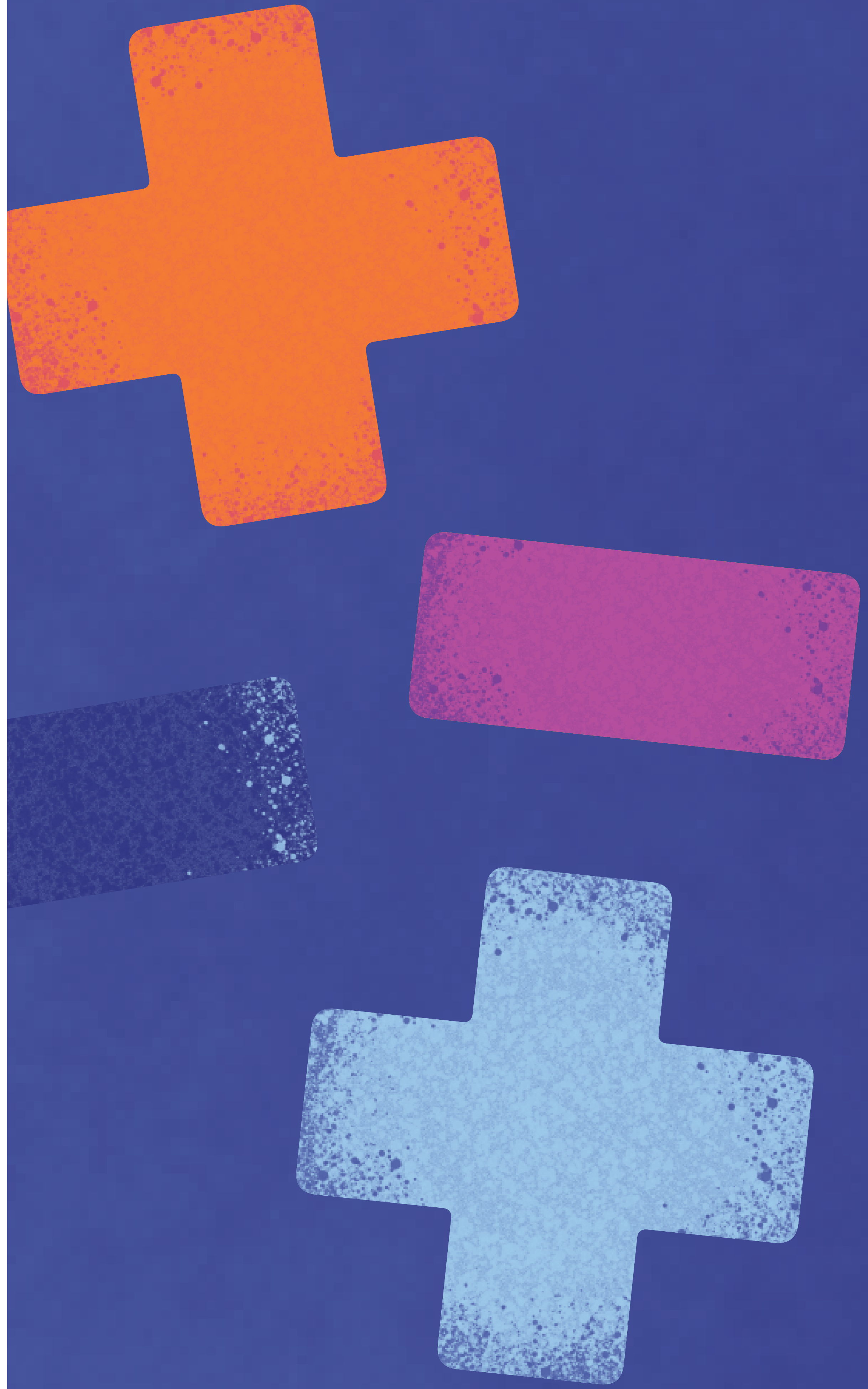




Another Example is thinking about statistical models that look at punishment - either in the criminal justice system or in educational discipline settings.

We have reasonably good data on what punishments are handing out - but not on what behaviors actually happened.

So we have isses that are very often accidentally biased.



Ways to think about fair

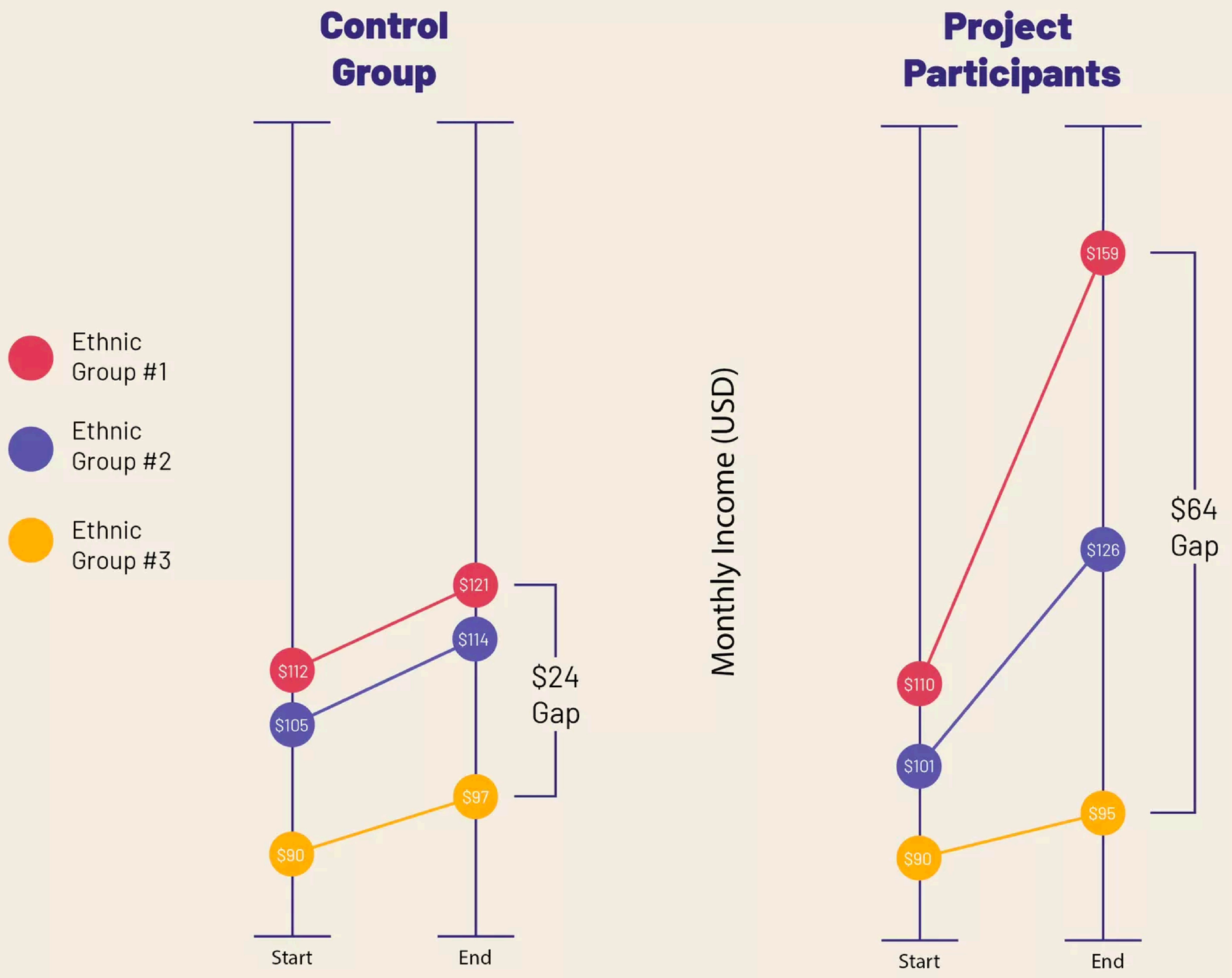
Equal False Negative Rates: the fraction of positives which are marked negative in each group agree.

Equal False Positive Rates: the fraction of negatives which are marked positive in each group agree.

Equal Positive Predictive Values: the fraction of those marked positive which are actually positive in each group agree.

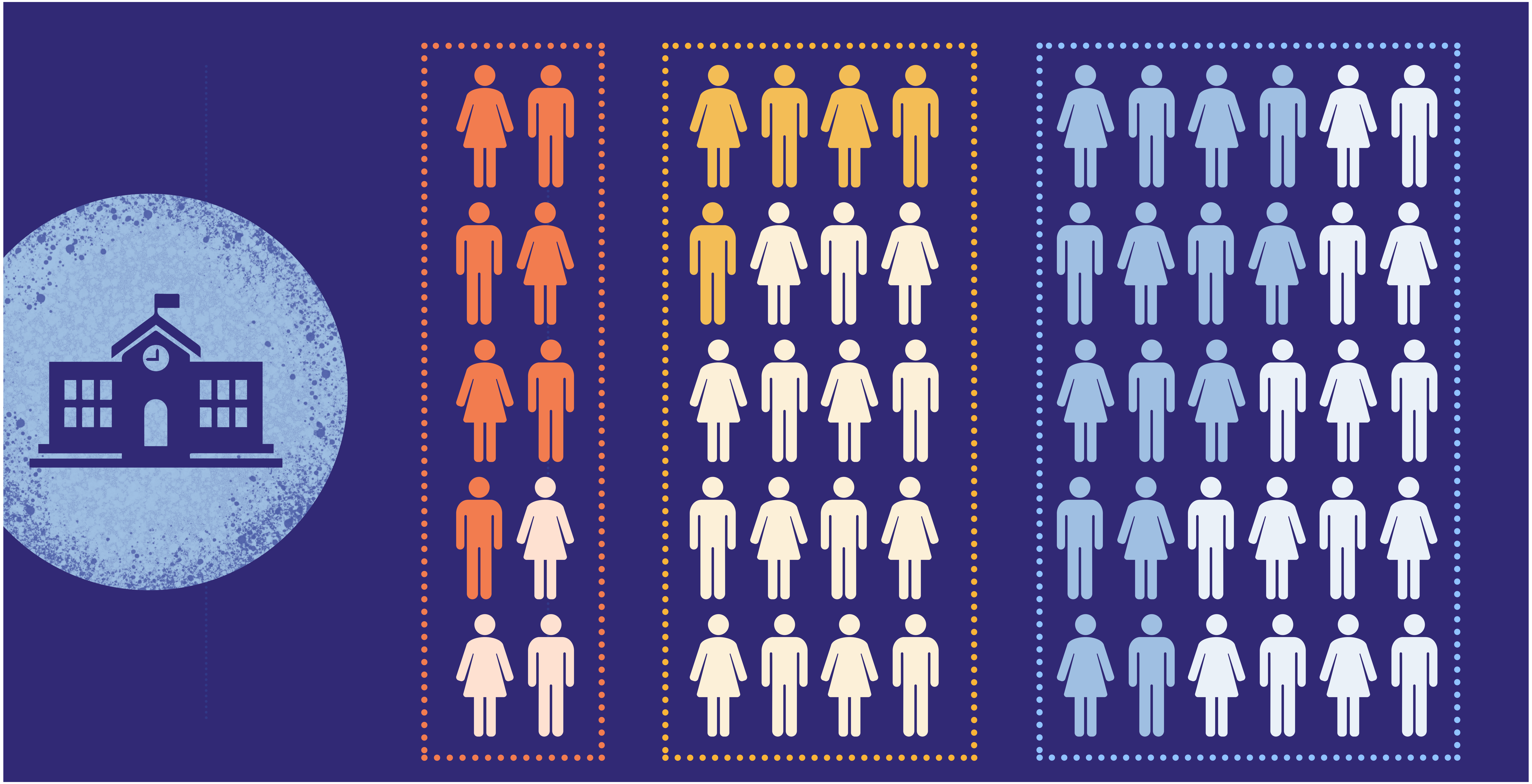
Statistical Parity (equal positive decision rates): the fraction marked positive in each group should agree.

Interpretation



Control Group

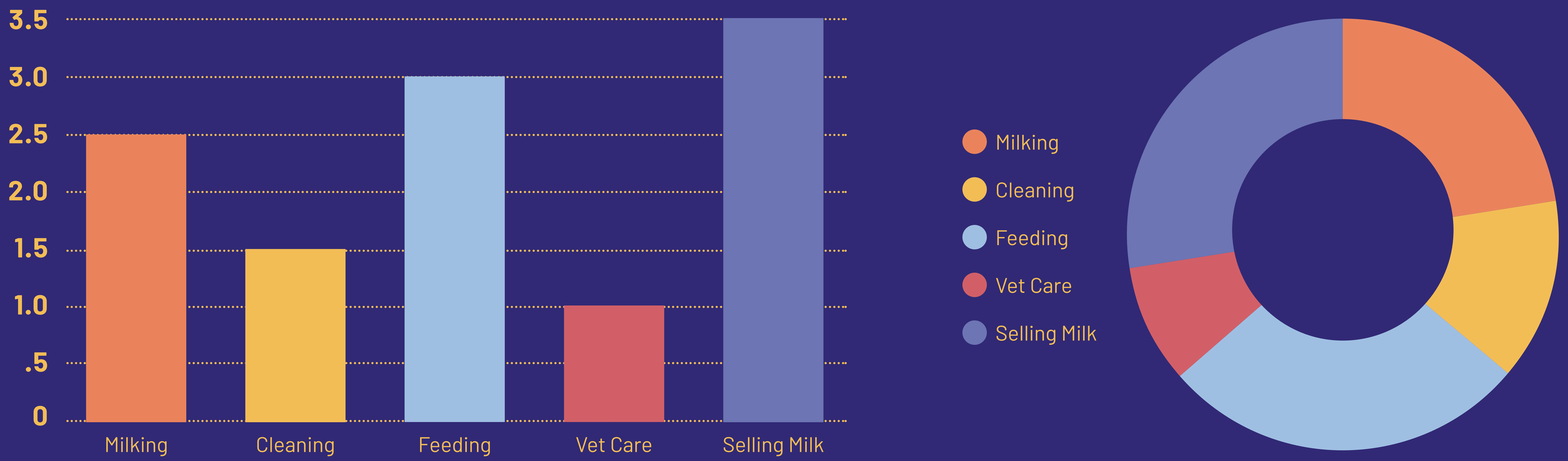
Project Participants



	Student Group 1	Student Group 2	Student Group 3
Number of Students	10	20	30
Count of Students Disciplined	7	5	15
Rate	75/100	25/100	50/100
Rate Relative to Student Group 2	3.0	1.0	2.0
Composition Index	27.3	18.2	54.5
Composition of Enrollment	16.7	33.3	50.0
Difference in Composition (Percentage Points)	10.6	-15.2	4.5
Relative Difference in Composition of Students Disciplines and Enrollment	63.6	-45.5	9.1

Communication & Distribution

Data Viz "best practices" are not culturally universaly.



TIME SPENT ON DAIRY ACTIVITIES PER DAY

Sources of bias can be identified in each step of the data life cycle.

Funding

Motivation

Project Design

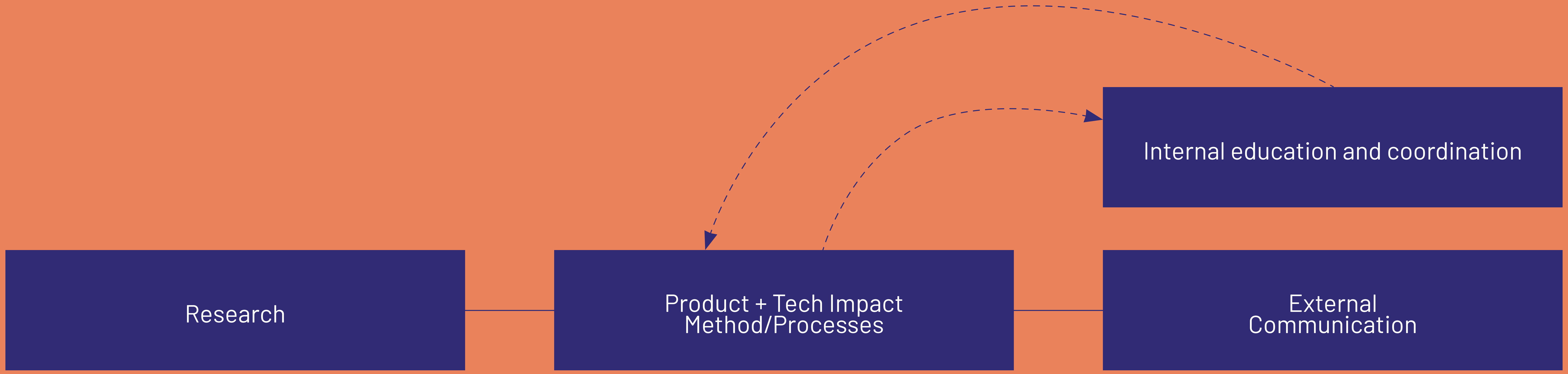
Data Collection & Sourcing

Analysis

Interpretation

Comunication & Distribution

Data Steward. You need a data steward that understand data ethics. Legal regulations, Internal practices and policies, External communication. Only about 1 in 5 companies have C-Suite involvement in data ethics - including privacy, consent, algorithmic accountability.


Internal education and coordination
Research
Product + Tech Impact Method/Processes
External Communication

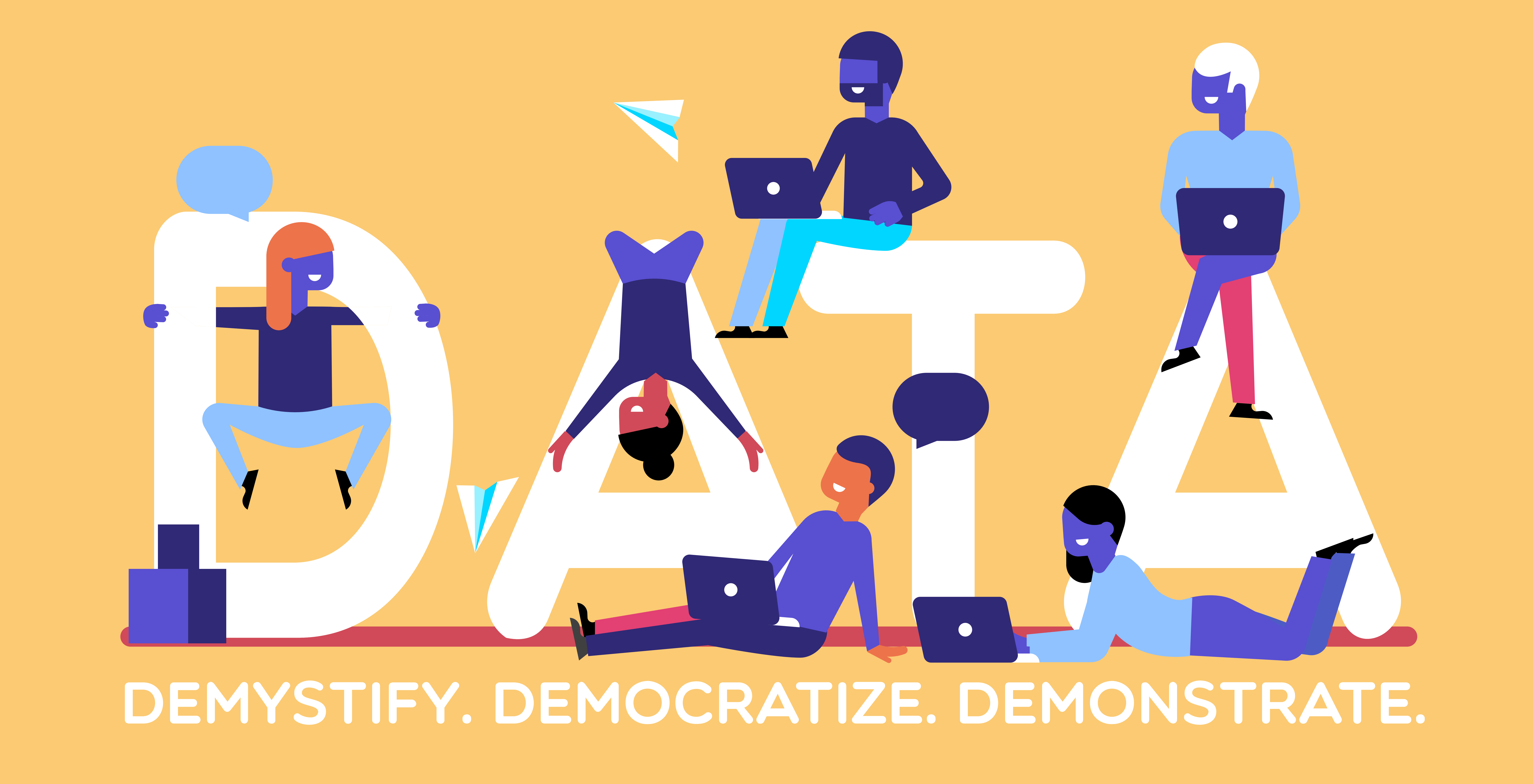
Thank you.

WeAllCount.com

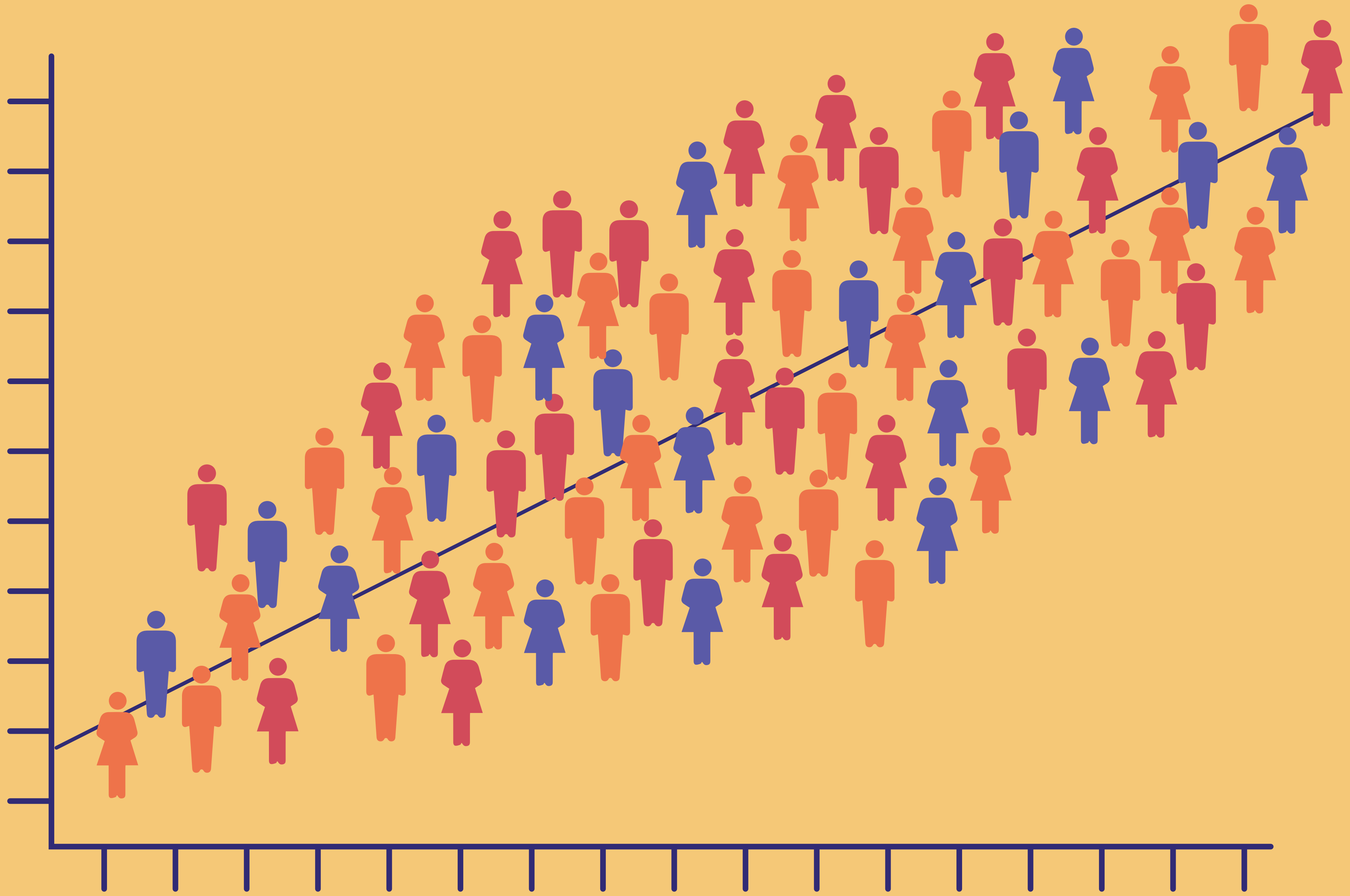
Heather Krause, PStat

heather@idatassist.com

@datassist



DATA
DEMYSTIFY. DEMOCRATIZE. DEMONSTRATE.



Based on data of previous generation.

Entirely different set of social and international circumstances.

Unfairly and mistakenly losing large potential customer base.

One of variables was immigration status and immigration class (refugee, family reunification, business, etc) .

This variable was associated with a negative coefficient.

We corrected their algorithm removing ethnic bias.

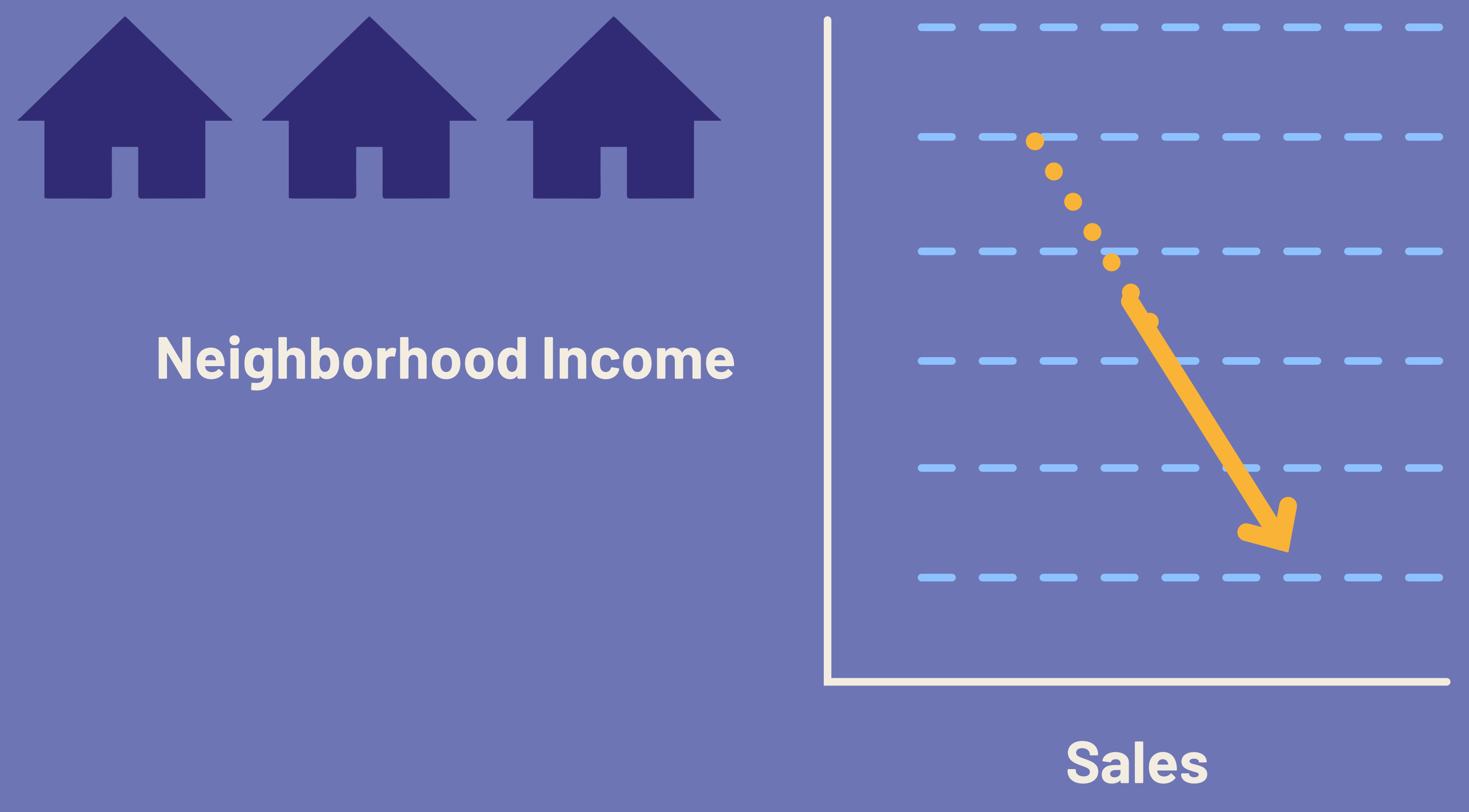
Newly acquired customers are currently creating same or higher value.

Worth 1.27 million

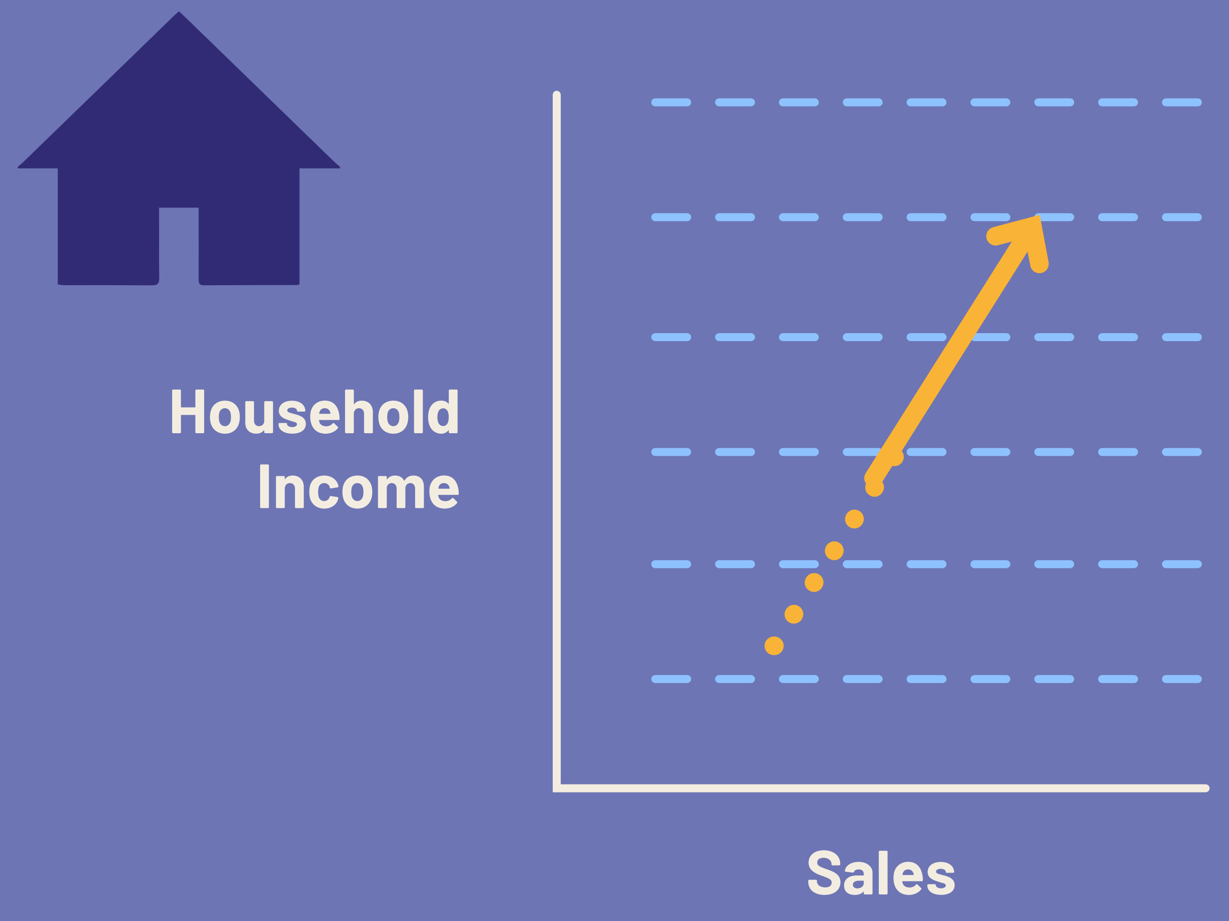
Story of **food delivery platform** I worked with.

Used **average neighbourhood house** cost as primary driver for customer acquisition.

Failure.



Negative correlation between **neighbourhood average income** and likelihood to buy the subscription.



Positive correlation between **individual household income** and likelihood to buy the subscription.